MEMOIRE DE FIN D'ETUDES

présenté pour l'obtention du diplôme d'Ingénieur Agronome

spécialisation : Amélioration des Plantes et Ingénierie végétale Méditerranéennes Et Tropicales (APIMET)

ANALYSE DE DIVERSITE ENTRE LES COMPARTIMENTS SAUVAGES ET CULTIVES CHEZ LE SORGHO : IDENTIFICATION DES GENES IMPLIQUES DANS LE PROCESSUS DE DOMESTICATION

PAR

Emmanuel RECLUS

Année de soutenance : 2012

Organisme d'accueil : UMR AGAP (Amélioration génétique et adaptation des plantes)

MEMOIRE DE FIN D'ETUDES

présenté pour l'obtention du diplôme d'Ingénieur Agronome

spécialisation : Amélioration des Plantes et Ingénierie végétale Méditerranéennes Et Tropicales (APIMET)

ANALYSE DE DIVERSITE ENTRE LES COMPARTIMENTS SAUVAGES ET CULTIVES CHEZ LE SORGHO : IDENTIFICATION DES GENES IMPLIQUES DANS LE PROCESSUS DE DOMESTICATION

PAR

Emmanuel RECLUS

Mémoire préparé sous la direction de :
Dominique THIS
Présenté le : 20/09/2012
Devant le Jury :
- Yves VIGOUROUX
- Vincent RANWEZ

Organisme d'accueil : UMR AGAP (Amélioration génétique et adaptation des plantes)

Maître de stage : David POT

Remerciements

J’adresse mes plus vifs remerciements à David POT, mon encadrant. Son écoute, sa patience et son soutien ont été infinis durant cette période de stage. Il m’est impossible d’exprimer une reconnaissance à la hauteur de sa considération et de sa disponibilité. Par la même occasion quelques petites pensées à Madame Plizz, Virginia et Jackie Sardou.

Je remercie également Dominique THIS, mon tuteur de stage pour son soutien et les nombreuses clés qu’elle m’a apporté pour perfectionner ma démarche scientifique.

Mes plus sincères remerciements à Jacques DAVID pour son œil bienveillant, à Sylvain GLEMIN et Benoit NABHOLZ pour leurs réponses, leurs conseils et nombreux apports scientifiques.

D’innombrables « merci » à Stéphane DE MITA pour la mise à disposition d’Egglib et surtout pour le service après-vente rapide, efficace, précis et agréable.

Que Monique DEU, Claire BILLOT et Jean-François RAMI trouvent ici l’expression de ma reconnaissance pour leur écoute, leurs conseils et leurs questions constructives.

Mes remerciements également à Letizia CAMUS-KULANDAIVELU pour nous avoir aidé à fixer nos idées sur les softs, à Sandy CONTRERAS pour toute son aide en programmation ainsi qu’à Nathalie CHANTRET pour ses coups de pouce en python.

J’adresse encore mes plus vifs remerciements à Felix HOMA, Gautier SARAH et Bertrand PITOLLAT qui ont consacré de nombreuses heures pour l’avancée des opérations. Leur patience a été soumise à rude épreuve, qu’ils trouvent ici toute l’expression de ma reconnaissance. A Gaëtan DROC et Alexis DEREEPER également qui ont toujours fait le maximum pour répondre à mes questions et mettre à ma disposition des outils performants.

A Angélique BERGER pour son soutien tout au long du stage et jusque dans les derniers jours ainsi qu’à Christian-Jacques ETIENNE pour toutes les discussions que nous avons eu et les croissants chaque semaine !

A Laura CHAPERON, Yan HOLTZ et Sophia HENRY sans qui mon moral n’aurait pas été si régulièrement positif et à tous ceux que j’oublie, à tort.

A mes très chers parents.

A Amandine.

Glossaire

(*) Merci de se référer au glossaire.

Abyss Cap3 Cap3 : Abyss est un assembleur *de novo* pour les reads courts et longs génomes (> 100 Mb). Cap3 est un autre assembleur capable de prendre en charge des séquences plus longues. La combinaison Abyss Cap3 Cap3 permet d'obtenir le meilleur consensus par réinjection des singlets et réassemblage des contig.

Basic Local Alignment and Search Tool (Blast) : est une méthode de recherche heuristique utilisée en bio-informatique permettant de trouver les régions similaires entre deux ou plusieurs séquences de nucléotides ou d'acides aminés. Ce programme permet de retrouver rapidement dans des bases de données, les séquences ayant des zones de similitude avec une séquence donnée (introduite par l'utilisateur) (Gibson, Muse *et al.*, 2004)

Cluster : groupement de séquences d'après l'homologie des séries de bases qui les composent.

Contig: Ensemble de fragments d'ADN clonés chevauchants pouvant être assemblés pour représenter une région définie du chromosome ou du génome duquel ils ont été obtenus. La définition des contigs est une étape nécessaire pour l'assemblage de séquences entières d'un génome (FAO, 2012).

Doublons optiques : séquences appartenant à un cluster mais identifiés lors du mapping comme appartenant à plusieurs clusters. Ils peuvent être identifiés sans alignements, simplement avec les coordonnées des séquences (lh3, 2012).

Fitness : la fitness d'un organisme (et donc d'une population) se définie par sa propriété à survivre ainsi que par sa fréquence de reproduction (taux moyen de descendants par unité de temps).

Indels : abréviation qui combine insertion et délétion. Dans l'exemple ci-dessous l'individu 1 possède une insertion de 9 bases par rapport à l'individu 2.

Individu 1 : ATCGTGACGTTGATCGTGCTAGTAACGTGACCAGT

Individu 2 : ATC−TGACGTT−−−−−−−−−−−−AACGTGACCAGT

Mapping, mapper : le mapping consiste à aligner les séquence sur une référence et ainsi à les assembler entre elles dans la mesure du possible.

Mismatch : est l'absence d'homologie entre deux bases de même position lors de l'alignement de deux séquences.

Outgroup : organisme ou groupe d'organismes qui sert de groupe de référence pour la détermination de la relation évolutive entre trois ou plusieurs groupes d'organismes monophylétiques.

Outlier(s) : gène(s) présentant un patron de diversité divergent des attendus neutre.

Read : séquence élémentaire issu du séquençage.

Score phred : score de qualité attribué à chacune des bases d'une séquence.

SNP : single nucleotide polymorphism. Variation ou substitution d'une seule paire de base au sein d'une séquence d'ADN entre deux individus.

Super contig : fragment non positionné dans le génome du sorgho mais identifié lors du séquençage comme en faisant parti.

Ti/Tv : Rapport entre le nombre de transitions et le nombre de transversions (Carr, 2010).

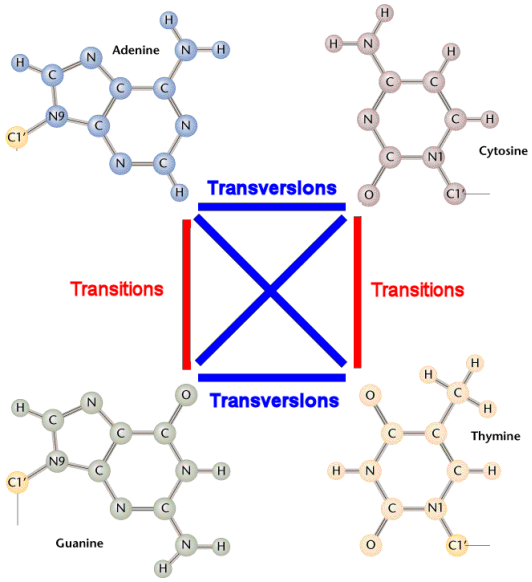


Table des abréviations

ADN (DNA) : acide désoxyribonucléique

ARF : ADP-ribosylation Factor

ARN (RNA) : acide ribonucléique

b :base(s)

Blast : Basic Local Alignment and Search Tool

d : date du goulot d'étranglement

D : durée du goulot d'étranglement

D_C : D de Tajima du compartiment cultivé

D_S : D de Tajima du compartiment sauvage

EggLib : Evolutionary Genetics and Genomics Library

f : force du goulot d'étranglement

GATK : Genome analysis toolkit

He : hétérozygotie

Icrisat : International Crops Research Institute for the Semi-Arid Tropics

Ind. : Individu

lseff : nombre de sites analysés

m : moyenne

Mb : méga base

Max : maximum

m_i : flux migratoire bidirectionnel

min : minimum

μ : taux de mutation par génération

MRCA : most recent common ancestor

N_0 : taille de la population sauvage

N_1 : taille de la population cultivée

N_b : taille de la population sauvage ayant subi le goulot d'étranglement

NA : donnée manquante

nseff : Nombre moyen de séquences analysées par sites analysés

pb : paire(s) de bases

π : diversité nucléotidique

Q1 : 1^{er} quartile (25% inférieurs des données)

Q3 : $3^{ème}$ quartile (25% supérieur des données)

sim : similarité

SSR : Simple Sequence Repeat

σ : écart-type

θ : estimateur de diversité

Table des figures

Table des tableaux

Introduction

Pour déchiffrer les bases génétiques de l'adaptation des plantes à l'état sauvage et dans les contextes agronomiques actuels, il est nécessaire de mieux comprendre comment la domestication a affecté leur diversité par une modification globale de leur génome. Une meilleure connaissance des phénomènes naturels et induits par l'homme qui participent à ce modelage permet de proposer des scénarios démographiques relatant l'histoire de vie des plantes et d'identifier ainsi des gènes impliqués dans ces processus qui peuvent par la suite être réintégrés dans les programmes de sélection.

Jusqu'à présent le faible nombre de marqueurs disponibles chez le sorgho ne fournissait pas une image assez fidèle de l'histoire évolutive du sorgho. La mise en évidence des gènes contrôlant les caractères d'intérêt agronomiques restait donc complexe sans la prise en compte de cette histoire. L'optimisation de l'identification de ces gènes, qui permettra ensuite une maximisation des gains génétiques dans les programmes de sélection, nécessite donc une compréhension plus fine de l'histoire évolutive du sorgho.

On se propose ici, grâce à une approche ABC, de définir le scénario d'évolution du sorgho le plus probable et grâce à une analyse de la diversité nucléotidique de la portion codante du génome d'un panel de 20 accessions représentatif de la diversité, de mettre en évidence des gènes de domestication et d'intérêt adaptatifs.

Dans la première partie du rapport, une synthèse bibliographique relative à la compréhension de la domestication et de l'histoire du sorgho est présentée. Les stratégies adoptées pour la détection et sélection des polymorphismes ainsi que la mise au point d'un modèle d'évolution du sorgho pour la mise en évidence des gènes cibles ont été abordées dans la seconde partie. Le rendu des résultats en troisième partie a ensuite été discuté. Cette discussion visant d'une part une comparaison de nos résultats obtenus avec les résultats déjà disponibles chez le sorgho et d'autres espèces et d'autre part une réflexion sur les limites des stratégies utilisées et des pistes d'optimisation.

1 L’étude de l’histoire des plantes pour mieux utiliser et conserver les ressources génétiques

La biodiversité des êtres vivants se manifeste par un panel de phénotypes extrêmement divers modelé par la dérive et les adaptations à des conditions de vie hétérogènes. Les ressources génétiques sont constituées par l’ensemble de ces variations qui sont « codées » au niveau du génome, les mutations étant la source de diversité. Bon nombre d’entre elles sont sélectivement neutres et ne confèrent aucun avantage sélectif et d’autres subissent la sélection en fonction de leur impact sur le phénotype. De ces ressources génétiques l’homme puise toute l’innovation pour l’amélioration et en modèle ainsi la diversité depuis la domestication des plantes (David, Loudet *et al.*, 2006). Les ressources génétiques sont donc au cœur d’enjeux mondiaux, tels que la sécurité alimentaire, la préservation de la biodiversité et l’adaptation de l’agriculture aux changements de l’environnement et de la demande sociétale.

L’optimisation de la gestion de la diversité et de l’efficacité de la sélection chez les plantes cultivées repose sur une meilleure connaissance des facteurs génétiques et des forces évolutives affectant la variabilité des caractères d’intérêt agronomique et adaptatif. Il s’agit donc de comprendre comment les gènes et les génomes ont été modelés par l'histoire, l'environnement et les sociétés.

Deux stratégies sont disponibles pour identifier les gènes d’intérêt agronomique et adaptatif (Ross-Ibarra, Morrell *et al.*, 2007). La plus largement utilisée, consiste, à partir de l’observation d’une variabilité phénotypique dans une population biparentale ou d’association, à identifier les régions du génome contrôlant cette variabilité par des approches de détection de QTL (du phénotype vers le gène).

Grâce à l’avènement des technologies de séquençage haut débit, une seconde approche indirecte consiste à mettre en évidence la diversité nucléotidique de l’ensemble du génome et ainsi identifier les régions portant des signatures des évènements de domestication ou d’adaptation à des contraintes naturelles. Il est ensuite possible de remonter jusqu’au gène impliqué dans le contrôle génétique de la variabilité de caractères soumis à la sélection naturelle ou artificielle.

2 L'étude de la domestication pour mieux comprendre l'évolution des génomes et identifier des gènes d'intérêt agronomique et adaptatif

La domestication a induit des changements importants au niveau des phénotypes et des génotypes. Ces transformations sont régies par des phénomènes identiques à ceux opérant en conditions naturelles – sélection, goulot d'étranglement, dérive, mutation, flux de gènes – mais sont guidées par l'homme et réduisent fortement la diversité génétique des compartiments domestiqués par rapport aux compartiments sauvages ancestraux. La domestication peut donc être considérée comme un modèle d'étude donnant accès à une meilleure connaissance des gènes qui y sont impliqués ou soumis à la sélection naturelle (Ross-Ibarra, Morrell *et al.*, 2007). Il en découle des apports en biologie évolutive et en biologie de la conservation ainsi qu'une meilleure compréhension des forces évolutives et de leurs interactions. Ces informations facilitent la mise au point de stratégies de gestion pour une conservation et valorisation des ressources génétiques plus durable et mieux adaptée aux besoins des populations.

Les études réalisées sur le riz, le maïs, le mil, le blé, ou l'orge constituent une source d'information intéressante pour mieux comprendre ces phénomènes chez le sorgho. Grâce à la synténie qui existe entre ces espèces il devient possible de cibler les analyses de diversité sur un compartiment de gènes sensible aux effets de la domestication.

Plusieurs études ont déjà permis de mettre en évidence des différences phénotypiques, entre individus cultivés et sauvages, qui sont sous le contrôle direct de gènes majeurs. A titre d'exemple, Doebley et al. en (1995) isole un QTL sur le chromosome 3 du maïs en grande partie responsable de ses différences morphologiques avec son ancêtre sauvage le téosinte et parvient à le cloner 2 ans plus tard (Doebley, Stec *et al.*, 1997). Il en est de même chez le riz avec le gène APETALA2 (AP2) dont le rôle est majeur dans le développement de la fleur (Gurian-Sherman, 2009). Une synthèse bibliographique des gènes impliqués dans le phénomène de domestication chez d'autres espèces de céréales a été effectuée il sera intéressant de tester si les gènes identifiés chez le sorgho comme portant des signatures de domestication correspondent à des orthologues des gènes identifiés chez d'autres espèces de céréales.

3 Le sorgho (*Sorghum bicolor ssp bicolor*), une céréale multi-usage

Préférentiellement autogame, le sorgho appartient à la famille des Poacées, c'est la 5ème céréale la plus importante en terme de production de grain et de surface de plantation (60 millions de tonnes sur 44 millions d'hectares) (Casa, Mitchell *et al.*, 2005; Zheng, Guo *et al.*, 2011). Le sorgho cultivé (*Sorghum bicolor* L. Moench) fait partie de la nourriture de base de plus de 500 millions de personnes à travers le monde (Frere, Prentis *et al.*, 2011). Son génome aujourd'hui disponible (Paterson, Bowers *et al.*, 2009), sa résistance aux stress abiotiques (sécheresse, chaleur et salinité), son métabolisme de type C4 et la très large diversité génétique de l'espèce (Casa, Pressoir *et al.*, 2008; Wang, Roe *et al.*, 2010) en font une des cibles les plus intéressantes pour répondre aux défis de la sécurité alimentaire et de l'accès à l'énergie.

3.1 Diversité, domestication et évolution du sorgho

Les trois sous espèces de sorgho (*Sorghum bicolor*) (fig. 1), qui comprennent les sorghos cultivés ainsi que leurs apparentés sauvages, sont : *S. bicolor* (L.) Moench (sorghos cultivés), *S. bicolor ssp. Verticilliflorum* (Steud.) Piper (sorgho sauvages) et *S. bicolor ssp. drummondii* (steud.) de Wet (hybrides stables).

Bien que les contraintes climatiques et d'ordre écologique affectent la distribution des différentes espèces, la diversité des sorghos sauvages et cultivés semble aujourd'hui être plus fortement liée à des facteurs géographiques et humains pour certaines régions d'Afrique plutôt que climatiques (Deu, Sagnard *et al.*, 2008; Mutegi, Sagnard *et al.*, 2011). Les sorghos cultivés présentent également une forte structuration en fonction de la race et de leur origine (Deu, Rattunde *et al.*, 2006) moins évidente chez les sorghos sauvages étant donné les difficultés rencontrées pour leur classification et les flux de gènes qui existent entre les compartiments sauvages et cultivés (Muraya, Mutegi *et al.*, 2011).

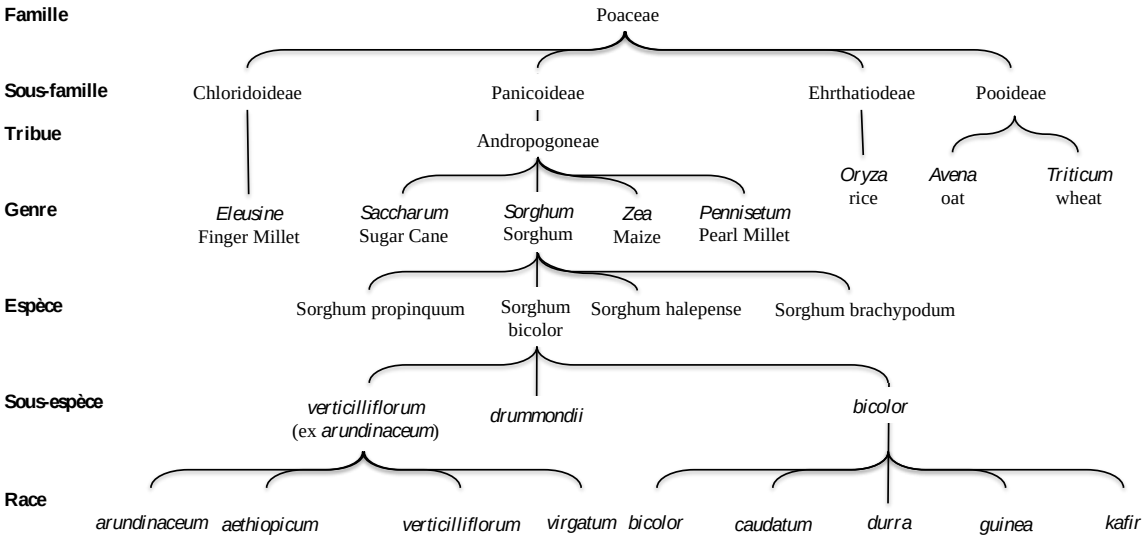


Figure 1 : Arbre taxonomique des Poaceae destiné à mettre en évidence la taxonomie du genre Sorghum (Trouche et Chantereau, 2009)

3.1.1 Les sorgho sauvages

Parmi les sorghos sauvages, il existe 4 races principales : *aethiopicum*, *virgatum*, *arundinaceum* et *verticilliflorum*. A la différence des sorghos domestiqués, ils ont conservé des traits qui confèrent une meilleure fitness* dans le milieu naturel comme par exemple la déhiscence précoce des graines, leur dormance et leur longévité. Leur durée de germination faible et le tallage sont des avantages sélectifs en situation de compétition (Sahoo, Schmidt *et al.*, 2010).

On peut discriminer les sorghos sauvages par la morphologie de leur inflorescence et leur distribution géographique (fig. 2 et 3) (de Wet et Harlan, 1971). *S. virgatum* possède de longues inflorescences étroites. Sa répartition s'étend de l'Egypte au Soudan tout comme celle de *S. aethiopicum* qui possède en revanche de petites inflorescences contractées. Les panicules de *S. verticilliflorum* sont larges et lâches. Il est le plus largement réparti dans la moitié sud de l'Afrique et entre le tropique du cancer et l'équateur. Enfin, *S. arundinaceum* se développe autour du golfe de Guinée jusqu'à la frontière du Soudan. Ses inflorescences ouvertes sont les plus imposantes par leurs dimensions. Plusieurs autres espèces sauvages *S. almum* Parodi, *S. purpureoserium* (Hochst. Ex A. Rich) Asch. & Schweinf, *S. halepense* (L.) Pers, et *S. versicolor* Andersson sont aussi identifiées cependant, leur morphologie et leur aire naturelle ne sont pas bien définies (Deu, Rattunde *et al.*, 2006).

Cette classification non évidente est en partie due aux critères utilisés pour discriminer les différents types. La hauteur de la plante, la taille des feuilles et des panicules, sont des traits quantitatifs peu fiables car influencés par l'environnement. Une autre raison possible qui peut contribuer à ces problèmes de classification est la très forte variabilité au sein d'une même population dans laquelle on peut trouver plusieurs taxons (Muraya, de Villiers *et al.*, 2011). Le sorgho bien que préférentiellement autogame (Pulchérie Barro-Kondombo, Vom Brocke *et al.*, 2008) présente des taux d'allogamie de 20% chez la race guinea par exemple (Barro-Kondombo, Sagnard *et al.*, 2010). De plus, le taux d'allogamie chez les sauvages semble être supérieur aux cultivés (Mutegi, Sagnard *et al.*, 2011; Sagnard, Deu *et al.*, 2011). En effet, plusieurs études montrent que la présence de flux de gènes, dont la régulation est multifactorielle devient un facteur d'évolution similaire et tend donc à réduire la divergence entre les deux compartiments lorsque les populations sauvages et cultivés sont voisines (Mutegi, Sagnard *et al.*, 2011; Sagnard, Deu *et al.*, 2011).

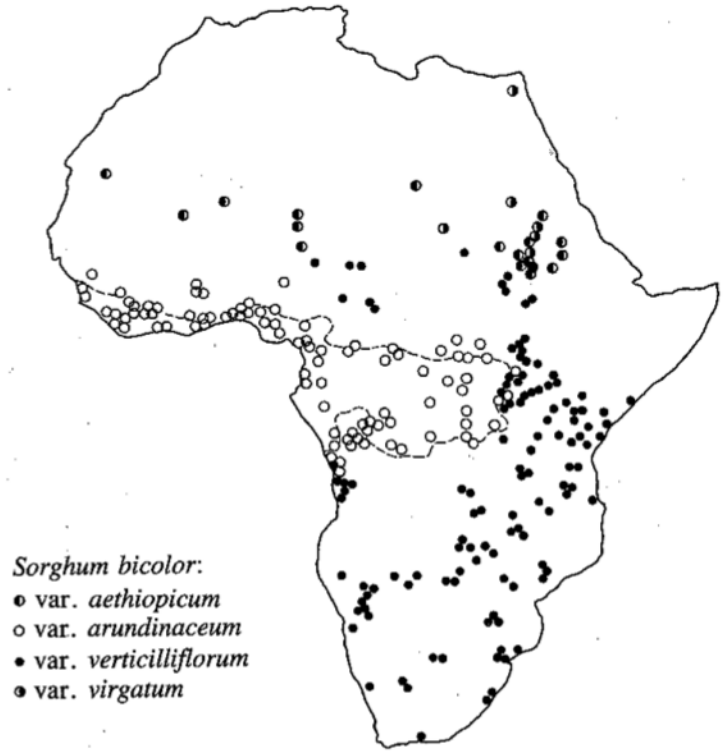


Figure 2 : Carte de répartition des sorghos sauvages en Afrique (de Wet, 1936)

Figure 3 : Photos de panicules de sorghos sauvages appartenant aux 4 races. En haut à gauche : S. aethiopicuum ; à droite : S. arundinaceum ; en bas à gauche : S. verticilliflorum ; à droite : S. virgatum. Lavalette, Cirad ; Berger, 2012.

Les individus sauvages et cultivés ont en effet la capacité de s'hybrider spontanément pour produire une descendance fertile (Tesso, Kapran *et al.*, 2008). La structuration dichotomique qui existe entre les deux compartiments étudiés à partir d'accessions clairement identifiées au sein des collections devient moins évidente lorsque les échantillons sont prélevés directement sur le terrain (Casa, Mitchell *et al.*, 2005).

Ceci facilite donc leur utilisation comme ressource génétique dans les programmes de sélection. Des mécanismes de résistance au Striga (adventice, parasite obligatoire des céréales) et une qualité d'amidon facilitant sa digestibilité ont notamment été trouvés dans les races sauvages, réservoir génétique important pour l'amélioration des cultivars (Muraya, Mutegi *et al.*, 2011).

3.1.2 Les sorghos cultivés

Il existe une grande diversité de variétés locales qui présentent une importante variabilité phénotypique. Elles sont cultivées dans des conditions agro-climatiques contrastées à l'aide d'une large diversité de pratiques culturales (Mutegi, Sagnard *et al.*, 2011). Plusieurs études (Grenier, Hamon *et al.*, 2001; Casa, Mitchell *et al.*, 2005; Deu, Rattunde *et al.*, 2006; Brown, Myles *et al.*, 2011; Bouchet, Pot *et al.*, 2012) aident à une meilleure compréhension de la structure génétique des sorghos cultivés à l'échelle mondiale (annexe 1). Bien que plusieurs hypothèses existent (Hamblin, Casa *et al.*, 2006; Sagnard, Deu *et al.*, 2011), on pense communément que le sorgho a été domestiqué en Afrique orientale (entre le lac Tchad et l'Ethiopie) il y a environ 3000 à 6000 ans[1] en donnant lieu aux premiers types de bicolor pour ensuite se répandre en Inde (env. 1500-1000 avant J.-C.), au Moyen-Orient (env. 900-700 avant J.-C.) et en Extrême-Orient (env. AD 400) (fig. 4). On parvient tout de même à les classer d'après la morphologie de la panicule (fig. 5) et des épillets en cinq races majeures (*bicolor, caudatum, durra, guinea*, et *kafir*) et dix races intermédiaires résultant de la combinaison par paire des précédentes.

[1] D'après certaines sources, la domestication serait plus ancienne. Environ 8000 ans avant notre ère (Dahlberg, 1995).

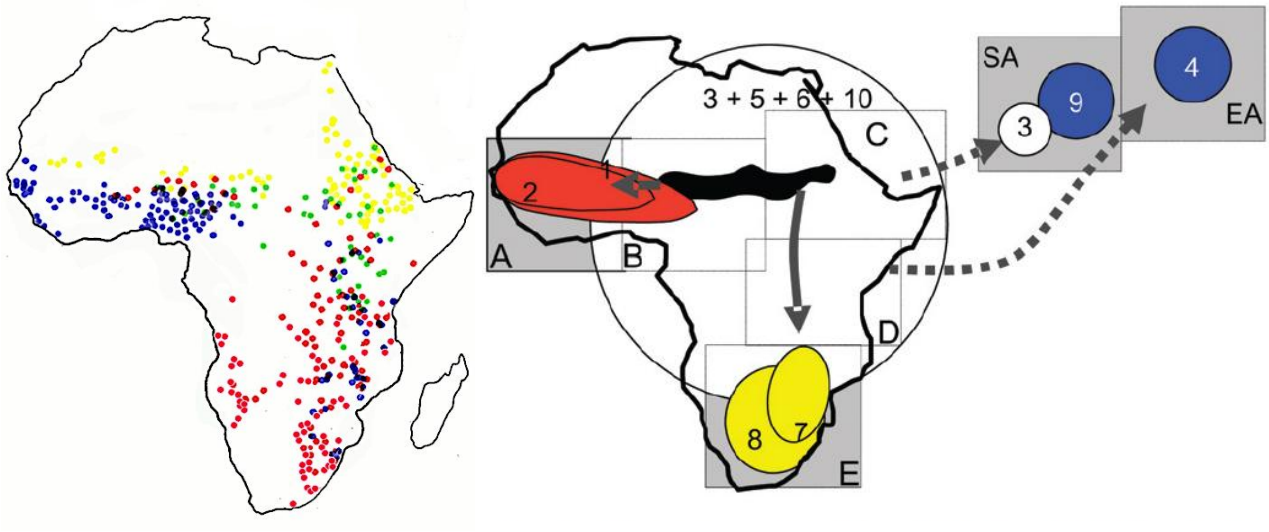


Figure 4 : A gauche : Carte de répartition des différentes races de sorghos cultivées en Afrique. En bleu : guinea, en vert : caudatum, en jaune : durra, en rouge. : kafir. A droite : Carte de l'Afrique présentant les centres initiaux de domestication (3, 5, 6, et 10 en noir) et les centres secondaires (1, 2, 4, 7, 8, 9). Les flèches représentent les principales migrations d'après Harlan (1995) (de Alencar Figueiredo, Calatayud et al., 2008)

Figure 5 : Photos de panicules de sorghos cultivés appartenant aux 5 races. En haut à gauche : Sorghum bicolor spp. bicolor race Bicolor ; au milieu : Sorghum bicolor spp. bicolor race Caudatum ; à droite : Sorghum bicolor spp. bicolor race Durra ; en bas à gauche : Sorghum bicolor spp. bicolor race Guinea ; à droite : Sorghum bicolor spp. bicolor race Kafir. Station d'expérimentation de Lavalette ; Pot, 2011.

La race *bicolor* présente une diversité élevée. Elle est considérée comme la race ayant la plus large distribution géographique et est aussi très importante par sa diversité d'usages (Deu, Rattunde *et al.*, 2006). Ensuite, les migrations, qui ont eu lieu il y a plus de 3000 ans, on donné naissance aux *guinea* qui se sont développés vers l'ouest et à la race *kafir* vers le sud (fig. 4). Cette dernière, contrairement à la race *guinea* divisée en trois groupes bien distincts et caractérisée par sa forte sensibilité à la photopériode, présente une diversité plus faible que les autres races (Dje, Heuertz *et al.*, 2000; Hamblin, Mitchell *et al.*, 2004; Casa, Mitchell *et al.*, 2005). Simultanément, la race caudatum s'est développée au niveau du centre d'origine et s'est répandue plus tard vers le sud et l'ouest. Les types *durra* sont prédominants en Asie du Sud et Afrique du Nord, il est difficile de savoir s'ils sont apparus d'abord en Afrique (Doggett, 1988) ou en Asie (Harlan, 1995). Ils sont caractérisés par des sorghos bien adaptés aux épisodes pluviaux en condition aride mais regroupent aussi des sorghos cultivés pendant la saison des pluies (Deu, Rattunde *et al.*, 2006).

3.2 Une histoire de vie complexe

Malgré l'importante variabilité phénotypique qui existe chez les sorghos cultivés, on constate au niveau moléculaire que la variabilité des sauvages est plus importante. Par exemple, l'étude de Casa *et al.* (2005) montre que les sorghos cultivés retiennent 86% de la diversité des sauvages. Les causes de cette baisse de diversité sont à la fois modernes et historiques. Dans le passé, la forte sélection pour les caractères d'intérêt agronomique liée aux épisodes de domestication a induit des goulots d'étranglements puissants réduisant ainsi la diversité du compartiment domestiqué par rapport à la population sauvage. D'après la littérature, plusieurs évènements de domestication seraient envisageables (Casa, Mitchell *et al.*, 2005; Hamblin, Casa *et al.*, 2006; Mutegi, Sagnard *et al.*, 2011; Sagnard, Deu *et al.*, 2011).

Aujourd'hui, la plupart des cultivars à rendement élevé sont issus de croisements de variétés génétiquement proches en faveur d'une meilleure adaptation et au détriment de la variabilité (Casa, Pressoir *et al.*, 2008). Suite à la forte réduction de diversité liée aux évènements de domestication, les sorghos cultivés ont subi une phase d'expansion grâce à la combinaison des migrations humaines et du commerce qui, dans de nombreux cas, ont permis aux plantes cultivées de se propager loin de leur centre d'origine (Purugganan et Fuller, 2009). L'évolution de ces populations domestiquées et leur adaptation à de nouvelles zones géographiques se sont accompagnées d'introgressions multifactorielles

modulées par la divergence (phénologie etc.), les pratiques agricoles, les échanges et la proximité des compartiments sauvages et cultivés.

3.3 Identification des gènes d'intérêt à l'aide de scénarios neutres

Plusieurs études rapportent la complexité de la détection de signatures de sélection sur des gènes candidats chez le sorgho cultivé (Hamblin, Casa *et al.*, 2006; Frere, Prentis *et al.*, 2011). Ces difficultés sont dues à de multiples facteurs – structuration de la population, goulots d'étranglement, introgressions, expansions, effet fondateur – qui influencent les paramètres démographiques (Frere, Prentis *et al.*, 2011).

Il est aujourd'hui possible de prendre en compte ces phénomènes démographiques. Les technologies haut débit – analyse des données de polymorphisme et de séquence – garantissent la production et l'obtention d'une information moléculaire pangénomique relativement fiable en un temps record. Avec cette capacité d'accès à de multiples loci et les méthodes de coalescence on peut estimer l'impact des forces évolutives et quantifier l'intensité et la date des goulots d'étranglement pour proposer des scénarios démographiques relatant l'histoire évolutive des plantes (Glemin, sd) qui permettent ensuite d'identifier des gènes soumis à la sélection.

Les objectifs de cette étude sont donc d'identifier les polymorphismes présents dans des échantillons représentatifs des compartiments cultivés et sauvages du sorgho afin de reconstruire son histoire évolutive et finalement de tenter d'identifier les gènes potentiellement impliqués dans le processus de sélection naturelle et de domestication. Dans la partie matériel et méthodes, j'exposerai les différentes étapes pour atteindre ces objectifs, de la détection et sélection des polymorphismes à la mise au point d'un modèle d'évolution du sorgho jusqu'à la mise en évidence des gènes impliqués dans la domestication ou apportant des avantages adaptatifs. Le rendu des résultats en troisième partie sera discuté et comparé avec les études précédentes avant de conclure.

4 Matériel et méthodes

4.1 Identification et caractérisation de polymorphismes sur le génome du sorgho

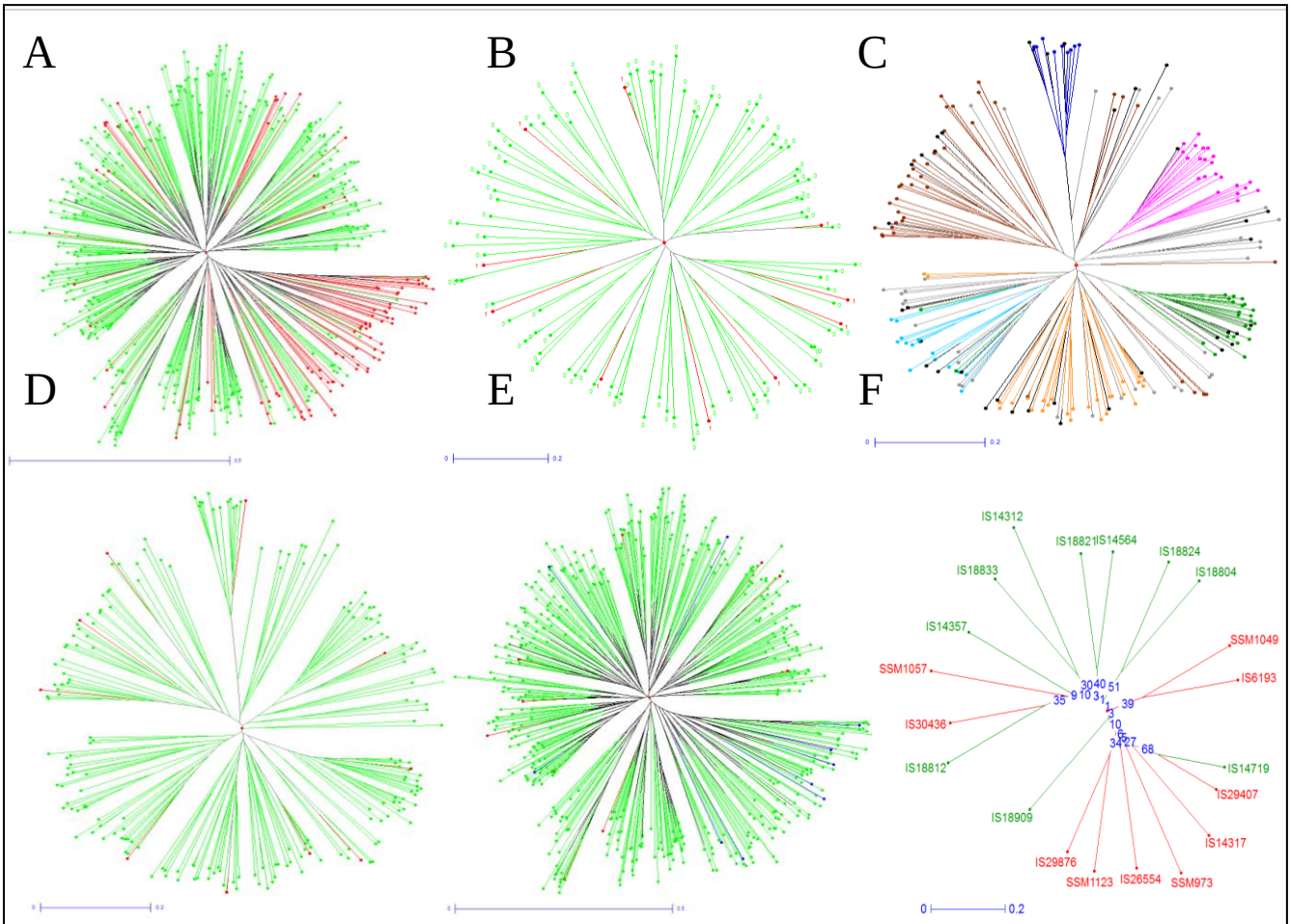
C'est une étape clé de la démarche. La sélection pertinente des polymorphismes permettra de dégager les résultats les plus proches possibles de la réalité. On considère que cette première phase s'étend de la sélection des génotypes jusqu'à la sélection des polymorphismes qui seront considérés pour les analyses évolutives.

4.1.1 Sélection des génotypes

Les 21 individus (10 sauvages, 10 cultivés et 1 outgroup) (tab. 1) étudiés ont été sélectionnés grâce aux résultats de différentes analyses de diversité réalisées à l'aide de plusieurs types de marqueurs (encadré 1). Cet échantillonnage vise à maximiser la couverture de la diversité (en termes de nombre d'allèles capturés et de nombre d'haplotypes). Le protocole d'échantillonnage est disponible en annexe 2.

Tableau 1 : Présentation des accessions utilisées dans cette étude

Ind.	Type	AccessionID	Sous espèce	Race	Pays
EB	Outgroup	Sorghum brachypodum	-	-	Australie
EC1	Cultivé	SSM1049	bicolor	B	Sénégal
EC2		IS29876	bicolor	C	Swaziland
EC3		IS30436	bicolor	C	Chine
EC4		SSM1123	bicolor	C	Niger
EC5		IS6193	bicolor	D	Inde
EC6		SSM973	bicolor	D	Sénégal
EC7		IS14317	bicolor	Gro	Swaziland
EC8		IS29407	bicolor	KC	Lesotho
EC9		SSM1057	bicolor	Gma	Sénégal
EC10		IS26554	bicolor	G	Benin
ES1	Sauvage	IS14564	verticilliflorum	aethiopicuum	Soudan
ES2		IS18821	verticilliflorum	aethiopicuum	Egypte
ES3		IS18909	verticilliflorum	aethiopicuum	Tchad
ES4		IS18824	verticilliflorum	arundinaceum	Côte d'Ivoire
ES5		IS18833	verticilliflorum	arundinaceum	Malawi
ES6		IS14312	verticilliflorum	verticilliflorum	Afrique du Sud
ES7		IS14357	verticilliflorum	verticilliflorum	Malawi
ES8		IS14719	verticilliflorum	verticilliflorum	Ethiopie
ES9		IS18804	verticilliflorum	virgatum	USA
ES10		IS18812	verticilliflorum	virgatum	Egypte



Encadré 1 : Arbres phylogénétiques (Neighbor Joining tree) illustrant le processus de sélection des accessions sauvages (AS) et cultivées (AC) utilisées. A : 413 accessions représentatives de la diversité mondiale (sauvages et cultivés) et 89 AS additionnelles. Elles ont été caractérisées avec 21 marqueurs microsatellites. Les AS sont indiquées en rouge, les AC en vert. La majeure partie des AS appartient à un même groupe de diversité (en bas à droite). B : 100 AS caractérisées avec 21 marqueurs microsatellites. Les accessions sélectionnées dans le cadre de cette analyse sont indiquées en rouge. C : 209 accessions cultivées issues de la Core Collection du CIRAD (Deu et al 2006) génotypées avec 41 marqueurs SSR. Les différents groupes génétiques identifiés par Bouchet et al (2012) sont indiqués en différentes couleurs (Marron : durra et bicolor d'Inde, caudatum et caudatum bicolor de Chine. Orange : caudatum et durra d'Afrique. Rose : guinea d'Afrique de l'Ouest. Bleu foncé : guinea margaritiferum d'Afrique de l'ouest. Bleu clair : guinea d'Afrique du Sud et d'Asie. Vert : kafir et kafir caudatum d'Afrique du Sud. Les accessions indiquées en gris n'ont pas pu être classées dans un de ces 6 groupes sur la base des analyses effectuées. Les accessions indiquées en noir n'ont pas été analysées par Bouchet et al (2012)). D : Localisation des accessions cultivées sélectionnées pour cette étude (en rouge) au sein des 209 accessions de la Core Collection du CIRAD. E : Localisation des 10 accessions cultivées (en rouge) et des 10 accessions sauvages (en bleu) sélectionnées pour cette étude au sein de 502 accessions représentatives de la diversité mondiale (Core Collection Cirad + Reference Set du GCP + 89 accessions sauvages additionnelles caractérisées avec 21 marqueurs microsatellite). F : Structure de la population analysée dans le cadre de cette étude, cette population comprend 10 accessions sauvages indiquées en vert et 10 accessions cultivées indiquées en rouge.

4.1.1 Extraction des ARN et séquençage

L'ARN issu des grains en cours de maturation, des feuilles et des fleurs (prélèvement des inflorescences au stade mi-anthèse) a été extrait grâce aux protocoles qui figurent en annexes 3-5 pour l'ensemble des individus. Des analyses à large échelle ont été effectuées chez d'autres espèces végétales (Sato, Antonio *et al.*, 2011) indiquant que ces trois tissus sont susceptibles de fournir un échantillonnage relativement complet du transcriptome. Les ARN ont été séquencés en paired-end (annexe 6) à l'aide de la technologie Genome Analyser de Solexa (HiSeq2000). Les mRNA des deux compartiments ont été séquencés en une ligne chacun et ceux de l'outgroup en une demie ligne. Les séquences de chaque génotype ont été marquées grâce à un adaptateur spécifique (annexe 7).

4.1.2 Nettoyage des séquences et mapping

Une vérification de la qualité a eu lieu sur les séquences brutes (format fastq) puis en fin de nettoyage à l'aide du programme FastqC (Babraham, 2012). Le nettoyage des séquences s'est déroulé en trois étapes : suppression des adaptateurs, filtration sur la qualité des bases (score phred min. 30) et la longueur des reads (min. 35) et comparaison des brins « forward » et « reverse » deux à deux. Les détails des paramètres se trouvent en annexe 8.

Une fois les séquences nettoyées et filtrées, le mapping consiste à positionner les séquences sur une référence grâce à l'homologie qui existe entre les deux séries de bases. Il est réalisé sur le génome et le transcriptome de Btx623[2]. Un nouveau nettoyage post-mapping a lieu lors duquel les doublons optiques*, les duplicatas de PCR (Rmdup_arcad) et séquences multi-mappées (Cleaner) sont éliminés avant l'identification des polymorphismes (annexe 9).

4.1.3 Détection et sélection des polymorphismes pour les analyses évolutives

Suite à la détection des polymorphismes via UnifiedGenotyper, le VariantFiltrationWalker caractérise les polymorphismes en fonction de leur qualité (PASS, SnpCluster, LowQual, Hard to validate). Ces caractéristiques sont explicitées dans le tableau 2. Certains d'entre eux cumulent différentes caractéristiques (exemple : Hard_to_validate;SnpCluster). Seuls

[2] Accession qui a été séquencée et annotée automatiquement : http://www.phytozome.net/dataUsagePolicy.php?org=Org_Sbicolor

les polymorphismes annotés « PASS » feront partie de la pré-sélection, à laquelle de nouveaux filtres, ont été appliqués pour la mise au point du jeu définitif.

Tableau 2 : Récapitulatif des caractéristiques attribuées aux polymorphismes par l'UnifiedGenotyper

Caractéristiques	Définition
HARD_TO_VALIDATE	Polymorphismes ne pouvant pas être classés en PASS car information insuffisante
LowQual	Score Phred trop faible (<40)
SnpCluster	3 polymorphismes ou plus sur 10 pb
PASS	Score Phred > 40

4.1.3.1 Sélection des polymorphismes et « recalibration »

Pour obtenir un set de polymorphismes contenant un minimum de faux positifs (artéfacts, problèmes d'alignement, présence de paralogues) nous avons appliqué les filtres suivants. Le premier consistait à recalibrer le jeu de polymorphismes obtenu sur la base de paramètres modulables propres à chaque polymorphisme et utilisés pour déterminer un score appelé VQSLOD (annexe 10), cette étape de recalibration correspondant à une évaluation de la qualité des polymorphismes détectés dans notre étude sur la base de jeux de polymorphismes déjà disponibles (dont la qualité est connue au préalable). Deux programmes implémentés dans GATK ont été utilisés pour la recalibration qui se base sur un jeu de SNP référence souvent issu de la littérature. Le VariantRecalibrator qui attribue le score de VQSLOD et ApplyRecalibration qui ne conserve que la tranche de SNP la plus pertinente en fonction de la distribution globale. Pour la mettre en œuvre, 4 jeux de SNP sources ont été utilisés (tab. 3).

La seconde méthode, inspirée de la recalibration, se base sur la similitude des Ti/Tv* avec un jeu de SNP. Il s'agit d'estimer la qualité d'un jeu de SNP en fonction de la valeur du ratio Ti/Tv. On se base sur l'hypothèse que les SNP faux positifs font diverger le Ti/Tv de sa vraie valeur (estimée à partir du jeu de SNP connus).

4.1.4 Mise au point du jeu de polymorphismes destinés aux analyses évolutives

Dans un premier temps, seuls les polymorphismes au niveau desquels une couverture de 8X chez au moins 8 individus a été observé ont été considérés. Dans un second temps, les polymorphismes hétérozygotes chez tous les individus (filtre 1), les polymorphismes hétérozygotes chez plus de 50% des individus cultivés (filtre 2) et les polymorphismes hétérozygotes chez plus de 70% des individus sauvages (filtre 3) ont été éliminés. Une estimation des taux de faux positifs et de faux négatifs a été tentée sur la base de données acquises par séquençage Sanger de 48 gènes dans 8 individus cultivés commun à ceux utilisées dans ce projet.

Enfin, préalablement aux analyses évolutives, seuls les alignements pour lesquels les informations de 6 accessions cultivées et 6 accessions sauvages ont été retenues soit 12 séquences de chaque population. Il est en outre important de mentionner que le faible effectif de génotypes analysé n'a pas permis d'avoir accès de façon précise à l'information haplotypique.

Tableau 3 : Récapitulatif des SNP sources utilisés pour la recalibration. Il est possible de moduler la confiance accordée aux SNP issus de 4 jeux (Sanger Cirad, US, Zheng, Nelson) par l'intermédiaire de 3 paramètres (known, training et truth). L'ensemble des SNP sources connus doit se trouver dans la catégorie « known ». Ce sont des SNP probablement vrais, ils ne seront pas utilisés dans le calcul de l'algorithme mais inclus dans le compte rendu final pour avoir une idée de leur distribution par rapport aux autres. Les SNP classés en tant que « training » seront utilisés dans l'algorithme pour le calcul du VQSLOD. Les SNP assignés « truth » permettent d'assigner la valeur de VQSLOD en deçà de laquelle les SNP sont éliminés. Il est également possible de moduler la qualité accordée aux SNP (paramètre « Qual »).

Origine	Nombre de SNP	Confiance	Known	Training	Truth	Qual.
Sanger Cirad + US + Communs Zheng-Neslson	2 176 82 193	100% Bonne	false	true	true	20.0
Zheng et al., 2011	1 112 672	+/- 99%	true	false	false	15.0
Nelson et al., 2011	283 528	82%	true	false	false	15.0
Sanger Cirad + US + Neslson	2 522	100%	true	false	false	20.0

4.1.5 Analyse des séquences non mappées

Suite aux mapping sur les références génome et transcriptome trois jeux de séquences non mappées ont été obtenus : les séquences qui ne mappent pas sur le génome (jeu n°1), les séquences qui mappent sur le génome mais pas sur le transcriptome (jeu n°2) et les séquences qui ne mappent pas sur le transcriptome (jeu n°3). Les jeux n°1 et n°3 sont directement extraits du fichier d'alignement global issu de l'assemblage des alignements individuels. Le jeu n°2 est obtenu grâce à un mapping sur génome du jeu n°3 duquel on extrait les séquences non mappées (jeu n°4).

Suite à un assemblage *de novo* des jeux n°1 et n°3 via Abyss Cap3 Cap3*, le mapping des mêmes séquences utilisées pour l'assemblage *de novo* sur ces nouvelles références permet de déterminer leur appartenance à un des 21 individus. Les homologies entre les contigs* issus de l'assemblage et les banques de biomolécules ont été recherchées par blastx*, afin de vérifier si ces contigs avaient une pertinence biologique (annexe 11).

4.2 Analyse des patrons de diversité nucléotidiques et identification des gènes affectés par les processus de sélection

4.2.1 Statistiques calculées sur les alignements

Les statistiques liées aux données moléculaires inter et intra groupes ont été estimées via la fonction *polymorphism* d'Egglib (De Mita et Siol, 2012) et confirmées par DnaSP (Rozas, Sanchez-DelBarrio *et al.*, 2003) et Arlequin (Excoffier, Laval *et al.*, 2005). Les statistiques et tests suivants ont été calculés sur les sites SNP uniquement, les indels* n'ont pas été considérés. Seules des analyses ne nécessitant pas l'information de phase des polymorphismes (haplotype) ont été utilisées.

Il existe plusieurs méthodes permettant l'estimation du paramètre mutationnel θ.

- L'estimateur de Watterson θ_w (Watterson, 1975) permet d'inférer le taux de mutation θ au sein d'une population. Avec N_e la taille efficace de la population et μ le taux de mutation par génération.

$$\hat{\theta}_w = 4N_e\mu$$

$\hat{\theta}_w$ tient compte de tous les allèles présents y compris ceux ayant une faible fréquence et plus particulièrement les mutants délétères par lesquels il peut être significativement affecté.

- Par contre, $\hat{\theta}_T$ l'estimateur du paramètre mutationnel de Tajima est estimé à partir des fréquences alléliques, les allèles de faible fréquence n'affectent donc pas beaucoup la valeur de $\hat{\theta}_T$. La diversité nucléotidique π est défini comme étant le nombre moyen de différences nucléotidiques par site entre deux séquences choisies aléatoirement dans un échantillon.

$$\hat{\theta}_T = \pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \eta_i \, i(n-i)$$

Avec η_i le nombre d'allèles dérivés ségrégés à une fréquence de *i/n* dans un échantillon de *n* chromosomes.

- Le test de Tajima D (Tajima, 1989) est basé sur la diversité moléculaire des échantillons. Le principe de ce test est de comparer les deux estimateurs décrits ci-dessus qui donnent des valeurs différentes de θ pour des populations ne vérifiant pas la neutralité. Soit *D* la différence de valeur entre les deux estimateurs de θ :

$$D = \frac{\hat{\theta}_T - \hat{\theta}_w}{\sqrt{V(\hat{\theta}_T - \hat{\theta}_w)}}$$

D = 0 dans le cas de la neutralité. Un D < 0 traduit un excès de variants de faible fréquence, on se retrouve dans le cas d'une sélection purifiante ou d'une expansion de population (après un goulot d'étranglement). Un D > 0 signifie un excès de variants de fréquence intermédiaire, c'est le cas lors d'une sélection balancée ou lorsque la population étudiée est structurée.

- Le Kst (Hudson, Slatkint *et al.*, 1992) est un test utilisé pour détecter la différenciation génétique d'une population subdivisée. Il se base sur le nombre de différences nucléotidiques entre des séquences issues de différents groupes et représente la proportion de la diversité nucléotidique totale attribuable à des différences génétiques entre les populations.

$$Kst = 1 - \left(\frac{K_s}{K_t}\right)$$

Avec Ks le nombre moyen pondéré des différences entre les sous-populations et Kt le nombre moyen de différences entre deux séquences de l'échantillon. Ces statistiques mesurent la proportion de la variation génétique qui est due à des différences génétiques entre les populations. Une valeur élevée signifie l'existence d'une subdivision de la

population et un Kst faible signifie qu'il n'y a pas de structure apparente dans la population.

4.2.2 Définition du scénario le plus probable d'évolution du sorgho

Pour détecter les gènes impliqués dans la domestication, il est nécessaire de reconstituer l'histoire évolutive de l'espèce. Il existe plusieurs possibilités pour définir les paramètres d'un scénario neutre. Nous avons choisi d'utiliser l'approche ABC (approximate Bayesian computation) (Pritchard et Rosenberg, 1999; Beaumont, Zhang *et al.*, 2002) implémentée dans le logiciel Egglib (De Mita et Siol, 2012). Cette méthode permet de calculer des lois a posteriori sans utiliser de vraisemblance mais en comparant directement les données simulées avec les données observées. D'une manière simplifiée la méthode ABC fait appel à un algorithme de type acceptation/rejet (fig. 6) et peut se décomposer en 3 étapes essentielles :

- La génération des paramètres à partir des lois a priori (theta, taille de la population cultivée, date, force et durée du goulot d'étranglement)

- La simulation des estimateurs de diversité avec ces paramètres (à l'aide du modèle) ;

- L'acceptation des paramètres simulés donnant lieu à la loi a posteriori si les données simulées sont proches des données observées ;

4.2.2.1 Paramètres démographiques et mutationnels du modèle DOM implémenté sous Egglib

On fait l'hypothèse (fig. 7) que la population cultivée subit une expansion une fois issue de la population sauvage suite à un goulot d'étranglement à la date *d* d'une durée *D* et d'une force *f*. On modélise le flux de gène qui existe entre les deux populations par *F*, le flux migratoire bidirectionnel estimé par $4N_0m_i$ avec m_i la proportion de population qui migre à chaque génération si on se place dans un référentiel discret. La force *f* du goulot est exprimée par le rapport entre la taille de la population pendant le goulot d'étranglement et celle de la population sauvage égale à 1.

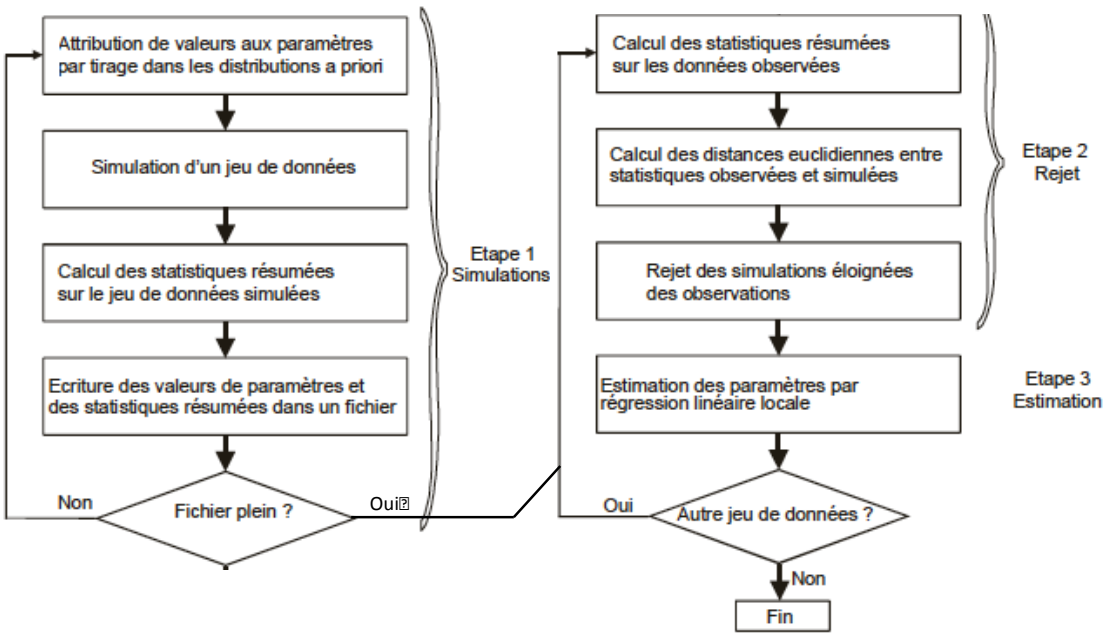


Figure 6 : Schéma des différentes étapes de l'approche ABC (d'après Cornuet et al., 2006)

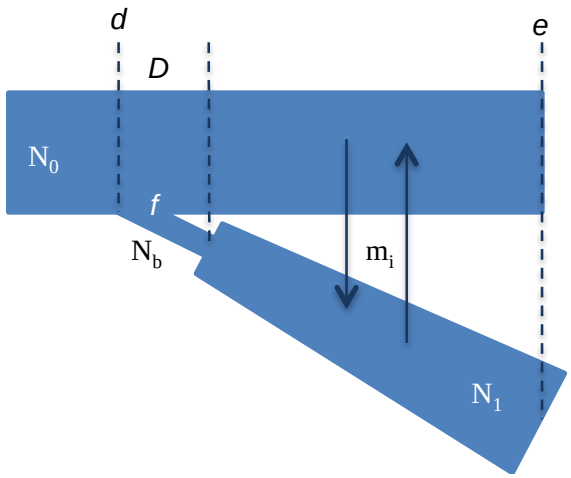


Figure 7 : Schéma représentant le modèle de domestication du sorgho. N_0 : taille de la population sauvage supposée constante ; N_1 : taille de la population cultivée ; N_b : taille de la population ayant subi le goulot d'étranglement ; d : date du goulot d'étranglement ; D : durée du goulot d'étranglement ; f : force du goulot d'étranglement ; m_i : flux migratoire bidirectionnel.

Le tableau 4 présente les distributions a priori des paramètres démographiques et mutationnels du modèle DOM implémenté dans le logiciel Egglib.

Au total, 1 100 000 simulations sont réalisées pour obtenir les distributions a posteriori, soit 100 000 pour chaque chromosome et le jeu de super contig*. L'ajustement des paramètres est réalisé grâce à la procédure de rejet simple. On ne conserve alors que 1% des données simulées, soit 11 000 points par paramètre (ditribution a posteriori).

La population sauvage est considéré comme égal à 1 et stable depuis l'époque de l'ancêtre commun le plus récent (N_0).

Le set de statistiques résumées pour estimer les distributions postérieures des paramètres comprend θ_W, π pour chaque population et le Kst.

4.2.3 Identification des gènes s'écartant du modèle neutre d'évolution

En raison de contraintes techniques et temporelles, deux stratégies ont été utilisées pour la recherche des outliers*. La distribution de l'indice de différenciation Kst entre les deux groupes d'accessions a été simulée 50 000 fois sous le modèle neutre obtenu via l'approche ABC. Afin de réduire le temps de simulation et en se basant sur le fait que les alignements présentent des propriétés redondantes en termes de longueurs de séquences et de nombre d'accessions disponibles pour les deux populations, les simulations n'ont pas été effectuées sur l'ensemble des 14 894 alignements mais sur un échantillon de 175 modèles de gènes (tab. 5) représentatifs de l'ensemble des gènes analysés. Cette distribution a été ensuite utilisée pour définir les intervalles de confiance du Kst qui ont permis d'identifier les loci présentant des différenciations extrêmes (en dehors des intervalles de confiance à 95, 99 et 99.9 %).

En ce qui concerne, les tests D de Tajima effectués au sein de chaque groupe d'accessions (Cultivés vs Sauvages), les distributions attendues sous le modèle identifié par l'approche ABC n'ont pas pu être obtenues. Dans ce cas, les 30 gènes présentant des valeurs extrêmes ont été retenus (15 avec les valeurs les plus basses et 15 avec les valeurs les plus élevées). La même stratégie a été adoptée pour le ratio de diversité entre le compartiment cultivé et sauvage.

Tableau 4 : Distributions a priori des paramètres démographiques et mutationnels : Loi uniforme [a ; b]. a et b correspondent aux bornes de l'intervalle des valeurs explorées

Paramètre	a	b
θ ($4N_0\mu$)	0,0	0,02
Taille t de la population cultivée (N_1/N_0)	0,0	1,0
Date d du goulot d'étranglement (normalisé par rapport à $4N_0$)	0,0	1,0
Durée D du goulot d'étranglement (normalisé par rapport à $4N_0$)	0,0	1,0
Force *k* du goulot d'étranglement (N_b/N_0)	0,0	1,0
Flux F migratoire bidirectionnel ($4N_0m_i$)	0,01	1,0

Tableau 5 : Propriétés des 175 gènes fictifs représentatifs des 14 894 alignements qui ont été utilisés pour générer les données simulées

	Longueur du gène (bp)	Nombre accessions sauvages	Nombre d'accessions cultivées
Min	100	6	6
Max	1900	10	10
Pas entre deux configurations	300	1	1
Nombre de configurations	7	5	5

5 Résultats

5.1 Identification des polymorphismes

5.1.1 Mapping sur les références « génome » et « transcriptome »

Au total, 611 250 316 séquences ont été générées par le séquençage. Après le 1er nettoyage, 481 412 914 ont été conservées (79%), elles se répartissent entre les groupes comme suit 57%, 30% et 13% pour les individus sauvages, cultivés et l'outgroup (fig. 8). Suite au nettoyage, on observe une perte globale d'environ 20% sur l'ensemble des individus (respectivement 23,9% et 16,2% chez les individus sauvages et cultivés).

La majorité (99%) des séquences mappées sur la référence du génome l'est aussi sur celle du transcriptome (soit 264 650 371 et 262 460 225 respectivement, correspondant à environ 43% des séquences de départ) (tab. 6). On constate cependant que 2 millions de séquences additionnelles sont cartographiées sur le génome et absentes sur le transcriptome.

Le nombre de séquences mappées chez les individus sauvages est plus important et plus hétérogène que chez les individus cultivés ($m=14878469$; $\sigma=5858787$ et $m=10318848$; $\sigma=2777861$ respectivement) (annexe 12).

5.1.2 Analyse des séquences non mappées

L'assemblage *de novo* des séquences non mappées sur transcriptome et génome donne lieu à 84 368 et 83 554 contig pour le transcriptome et génome (tab. 7).

A partir de l'assemblage *de novo* des jeux n°1 et 3 on constate que 60% des contigs assemblés sont annotés. L'analyse du jeu n°2 ainsi que la définition de l'origine de ces contigs n'ont pas été réalisées dans cette étude.

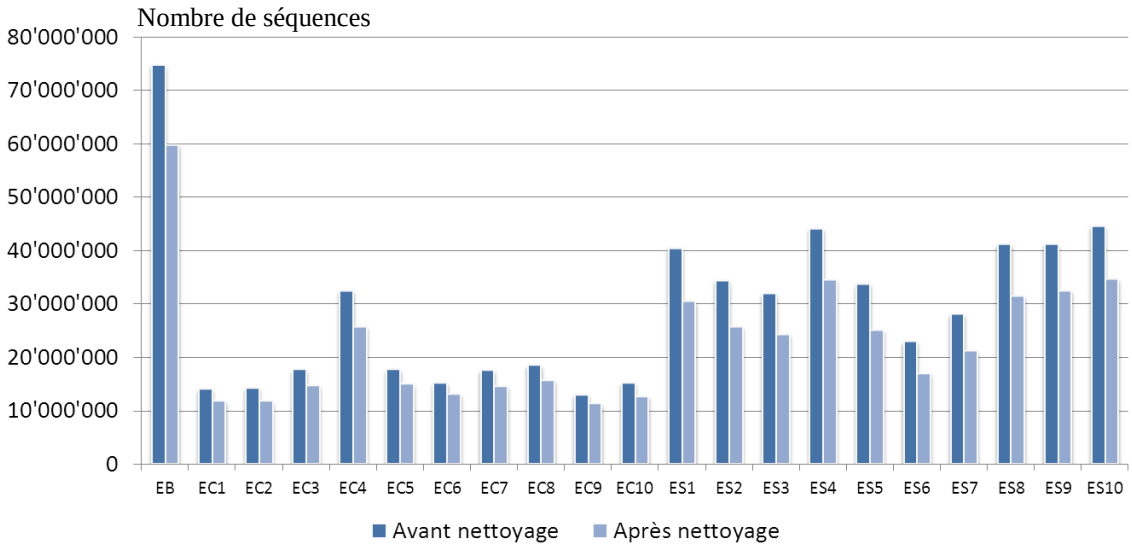


Figure 8 : Nombre de séquences obtenues par accession avant et après nettoyage

Tableau 6 : Nombre de séquences conservées après les différentes étapes jusqu'au mapping

	Transcriptome	**Génome**
Séquences après 1er nettoyage	481 412 914	
Nombre séquences mappées	262 460 225	264 650 371
Séquences éliminées après mapping	102 084 405	134 740 047
Nombre séquences non mappées	116 868 284	82 022 496

Tableau 7 : Bilan de l'analyse des séquences non mappées

	Transcriptome	**Génome**
Séquences non mappées	116 868 284	82 022 496
Contig assemblés	84 368	83 554
Singlets (2nd Cap3)	22 8783	22 6236
Contig annotés	50 175 (60%)	49 820 (60%)

5.1.3 Sélection du set de polymorphismes destinés aux analyses évolutives

5.1.3.1 Caractérisation de l'ensemble des polymorphismes détectés

Le tableau 8 présente les différentes catégories de polymorphismes détectés sur le transcriptome et le génome. La distribution de la qualité des polymorphismes issus du transcriptome paraît plus intéressante. Les polymorphismes qualifiés « PASS » y sont majoritaires alors qu'à l'issu d'un mapping sur le génome une partie d'entre eux est qualifiée de SnpCluster.

Le jeu de polymorphismes destiné aux analyses évolutives a été détecté grâce à la version 1.6 du GenomeAnalysisTK.jar de GATK (McKenna, Hanna *et al.*, 2010). Les filtres appliqués aux sites polymorphes candidats sont explicités dans le tableau 9. A partir des 1 018 393 (1 023 604 avec la version 1.4 du GenomeAnalysisTK.jar) polymorphismes identifiés sur le transcriptome, 501 009 ont été conservés pour les analyses évolutives.

Concernant les polymorphismes issus des individus cultivés, on s'attend au minimum à 85 % d'homozygotie. Pour ne pas être trop restrictif, seuls les polymorphismes hétérozygotes chez plus de 50% des individus ont été éliminés. Ceux issus des sauvages sont éliminés s'ils sont hétérozygotes chez plus de 70% des individus, sachant qu'on attend 50% d'individus hétérozygotes maximum.

Les filtres 1, 2 et 3 ont finalement mené à l'élimination de 13 297 polymorphismes provenant de 4395 gènes. Pour 4332 de ces gènes, des polymorphismes ont néanmoins été conservés dans le jeu final (175 110 en tout) et analysés dans le cadre de cette étude.

Tableau 8 : Caractéristiques des polymorphismes identifiés grâce au GenomeAnalysisTK.jar (version 1.4) de GATK. (a) : Pourcentage des polymorphismes présents sur le transcriptome (en %), (b) : Pourcentage des polymorphismes présents sur le génome (en %).

SNP et indels	Transcriptome	F_T [a]	Genome	F_G [b]
HARD_TO_VALIDATE;SnpCluster	5 362	0,52	12 964	0,58
HARD_TO_VALIDATE;LowQual;SnpCluster	1 621	0,16	3 436	0,15
HARD_TO_VALIDATE	16 506	1,61	13 437	0,60
LowQual;SnpCluster	55 862	5,46	183 795	8,17
HARD_TO_VALIDATE;LowQual	5 749	0,5	6 123	0,2
LowQual	128 763	12,58	373 060	16,59
SnpCluster	234 708	22,93	745 631	33,15
PASS	575 033	56,18	910 694	40,49
Total	**1 023 604**	**100,00**	**2 249 140**	**100,00**

Tableau 9 : Récapitulatif des filtres appliqués aux sites polymorphes candidats

	Polymorphismes éliminés	Total après application du filtre
Polymorphismes candidats (HARD_TO_VALIDATE;SnpCluster ; HARD_TO_VALIDATE ; SnpCluster et PASS)	- 192 194	826 199
Couverture ≥ 8X et information disponible pour au moins 8 individus	- 184 097	642 102
Présence de l'allèle en fréquence minoritaire chez au moins 1 individu respectant la couverture de 8X	- 126 046	516 056
Allèle de la référence présente au moins une fois dans le set de génotypes analysés dans ce projet (polymorphisme non spécifique de Btx623)	- 1 886	514 170
Homozygote chez au moins 1 individu (présence de l'allèle alternatif au moins une fois) (filtre 1)	- 1 937	512 233
Nombre d'individus hétérozygotes > 50% chez les individus cultivés (filtre 2)	- 9 550	502 683
Polymorphismes sans couverture chez les cultivés (impossible d'évaluer l'hétérozygotie)	+ 112	502 795
Nombre d'individus hétérozygotes > 70% chez les individus sauvages (filtre 3)	- 1810	500 985
Polymorphismes sans couverture chez les sauvages (impossible d'évaluer l'hétérozygotie)	+ 24	**501 009**

5.1.3.2 Premières information issues du jeu de polymorphismes

Dans le tableau 10, on constate que le compartiment cultivé présente beaucoup moins de polymorphismes spécifiques (25 919) que le compartiment sauvage (92 304) alors que le ratio entre le nombre de polymorphismes dans le compartiment cultivé et sauvage est de 0,61. Une plus grande diversité observée au sein du compartiment sauvage par rapport au cultivé peut s'expliquer par une réduction de la diversité au cours de l'évènement de domestication et par une probabilité de détection plus élevée de polymorphismes en raison du nombre de séquences plus important chez les sauvages. Les indices de fixation observés chez les individus cultivés et sauvages sont conformes aux attentes (hétérozygotie plus élevée chez les sauvages).

5.1.3.3 Estimation du taux de faux positifs et faux négatifs

Nous avons utilisé 309 SNP issus de 48 gènes (jeu Sanger) qui ont été séquencés chez 8 individus pour estimer la véracité de notre jeu de polymorphismes. Il apparaît que seulement 80 SNP sont communs aux deux jeux (585 polymorphismes détectés sur les séquences obtenues dans cette étude). Ce faible recouvrement des polymorphismes détectés s'explique, d'une part par une faible couverture des gènes analysés en Sanger au sein des séquences produites dans le cadre de cette étude. En effet, pour la majeure partie des gènes analysés une couverture inférieure à 2X pour l'ensemble des accession séquencées a été obtenue ce qui augmente donc de façon très significative la probabilité de non détection de polymorphismes (faux négatifs). D'autre part, cette étude a permis l'analyse de régions des gènes qui n'avaient pas été analysées en Sanger ce qui a donc induit la détection de nouveaux polymorphismes. Le séquençage en Sanger correspondait en effet à environ 50% des CDS des gènes.

5.1.4 Définition du jeu de séquences utilisé pour les analyses évolutives

En regard des 16 122 gènes pour lesquels des polymorphismes ont été mis en évidence, plusieurs milliers de gènes ne présentaient pas de polymorphisme. Ces derniers sont aussi importants à prendre en considération pour comprendre l'histoire évolutive du sorgho. Au total, 24 727 alignements ont donc été considérés sur lesquels deux filtres ont été appliqués. Le premier sur le nombre de données exploitables (annexe 13), le deuxième sur le nombre de séquences par compartiment (≥ 12) et le nombre de bases par séquences (≥ 100) (tab. 11). Les séquences de l'outgroup n'ont pas été utilisées.

Tableau 10 : Caractéristiques des 501 009 polymorphismes identifiés sur le transcriptome après filtration

Niveau d'analyse	Variable	Valeur
Global	Nombre de gènes polymorphes	16 122
	Nombre de sites polymorphes	501 009
	Nombre de SNP	462 054
	Nombre d'indels	38 955
	Ratio Ti/Tv	1.76
Intraspécifique/ Intragroupe	Nombre de polymorphismes dans *S. brachypodum*	53 894
	Nombre de polymorphismes spécifiques à *S. brachypodum*	50 933
	Nombre de polymorphismes dans le compartiment cultivé (EC)	105 707
	Nombre de polymorphismes spécifiques du compartiment cultivé (EC)	25 919
	Nombre de polymorphismes dans le compartiment sauvage (ES)	173 042
	Nombre de polymorphismes spécifiques du compartiment sauvage (ES)	92 304
	FIS moyen pour le compartiment cultivé	0.77
	FIS moyen pour le compartiment sauvage	0.23
Interspécifique / Intergroupes	Nombre de polymorphismes fixés entre *S. brachypodum* et les *S. bicolor*	250 604
	Nombre de polymorphismes fixés entre les cultivés et les sauvages	45
	Nombre de polymorphismes pour lesquels l'allèle en fréquence minoritaire est présent uniquement chez les hétérozygotes	114 626
	Nombre de polymorphismes détectés chez les cultivés (EC) pour lesquels l'allèle en fréquence minoritaire est présent uniquement chez les hétérozygotes	18 251
	Nombre de polymorphismes détectés chez les sauvages (ES) pour lesquels l'allèle en fréquence minoritaire est présent uniquement chez les hétérozygotes	80 459

Tableau 11 : Récapitulatif du nombre d'alignements validés et filtrés pour les analyses évolutives

Chrom	Validés	Filtrés sur DE[a]	Filtrés sur nS[b] et nB[b]	Total
Sb[1]	72	11	25	108
Sb01	2849	324	1219	4392
Sb02	1942	298	1048	3288
Sb03	2215	244	1057	3516
Sb04	1806	226	826	2858
Sb05	521	808	12	1341
Sb06	1353	189	664	2206
Sb07	955	158	537	1650
Sb08	702	131	479	1312
Sb09	1250	161	580	1991
Sb10	1229	178	658	2065
Total	14894	2728	7105	24727

(a) : DE : données exploitables ; (b) : nS : nombre de séquences ; (c) : nB : nombre de bases ; (1) : super contigs.

Au final 175 110 polymorphismes ont été considérés pour les analyses évolutives. Ils proviennent de 14 894 contigs parmi lesquels 14 150 sont polymorphes.

5.1.5 Estimateurs de la diversité nucléotidique et tests de neutralité sélective sur les compartiments cultivés et sauvages

θ_w (estimateur du taux de mutation de la population), la diversité nucléotidique (estimateur du taux de mutation de la population de Tajima) et l'hétérozygotie semblent plus élevés chez les individus sauvages que chez les cultivés (tab. 12). Un test statistique de type ANOVA serait nécessaire pour s'en assurer. La moyenne du Kst étant plutôt faible, elle nous informe de la présence de structure.

5.1.6 Définition du scénario d'évolution

5.1.6.1 Stabilité des postérieurs en fonction des nombres de simulations effectuées

Le temps d'analyse étant une contrainte majeure, nous avons défini pour quel nombre de simulations les paramètres se stabilisaient. Après avoir sélectionné 100 alignements contenant le plus grand nombre d'informations (nombre de sites exploitables et de séquences par population), nous avons appliqué 10 000, 100 000 et 1 million de simulations. Les distributions sont résumées en annexes 14 et 14bis.

Il s'avère que les distributions générées par 100 000 et 1 million de simulations sont relativement proches, il a donc été décidé de réaliser 100 000 simulations pour l'ensemble des gènes.

Tableau 12 : Récapitulatif de la distribution des estimateurs de diversité θ, π, D de Tajima, He et Kst sur 14894 gènes

	Sauvages	Cultivés	
θ	0	0	min
	0,003209645	0,002052869	m
	0,002908006	0,002471791	σ
	0,060048445	0,036834748	max
π	0	0	min
	0,003237768	0,002258203	m
	0,0035302	0,003054481	σ
	0,073841471	0,055503704	max
D	-2,46914362	-3,388793791	min
	-0,170404458	0,14685539	m
	1,135450012	1,017076156	σ
	3,117851345	3,231641274	max
	980	1936	NA
He	0	0	min
	0,566077249	0,477995964	m
	0,290695745	0,283076833	σ
	1	0,994736842	max
Kst	-0,077142857		min
	0,0402033		m
	0,058807109		σ
	1		max
	744		NA

Tableau 13 : Distributions a posteriori des paramètres démographiques et mutationnels obtenus à partir de 100 000 simulations avec le modèle DOM de Egglib

	Médiane	Q1	Q3
θ ($4N_0\mu$)	**0,00105**	0,00049	0,00170
Taille de la population cultivée (N_1/N_0)	**0,64600**	0,47090	0,76250
Date du goulot d'étranglement (normalisé par rapport à 4N0)	**0,02800**	0,01280	0,05130
Durée du goulot d'étranglement (normalisé par rapport à 4N0)	**0,02600**	0,01170	0,04800
Force du goulot d'étranglement (N_b/N_0)	**0,64023**	0,39680	0,84744
Flux migratoire bidirectionnel ($4N_0m_i$)	**0,53223**	0,35021	0,63306

5.1.6.2 Définition des valeurs les plus probables des différents paramètres du modèle

Les distributions a posteriori sont présentées dans le tableaux 13. Remarque : la valeur estimée des 6 paramètres appartient aux intervalles définis a priori.

Sur la base de ces résultats :

- Avec la valeur de θ obtenue sur les 14894 gènes il est possible de calculer la taille effective de la population sauvage (N_0). Si μ=1.0e-08 (Hamblin, Casa *et al.*, 2006) ; N_0 = θ/4μ = 1.05e-03/(4x1.0e-08) = 26 250.

- Sachant que N_0 = 26 250, la taille de la population cultivée N_1 est de 16 958.

- On estime de la date du goulot d'étranglement, sachant N_0 = 26 250, à 2 946 ans et sa durée à 2 735 ans correspondant à une date de domestication il y a 5681 années.

- Force du goulot d'étranglement égale à N_b/N_0 avec N_b la taille de la population pendant le goulot d'étranglement. Soit N_b = 0.640233x26 250 = 16 806. Et $k = N_b/D$ = 16 806/2 735 = 6.1 signifierait que la force du goulot d'étranglement est très élevée.

- Taux de migration bidirectionnel égal à $4N0m_i$. Soit m_i = F/(4x26 250) = 0,00000507

En fonction du taux de mutation considéré il apparaît des variations notoires dans les estimations. On propose de reprendre les taux de mutations du maïs dans le tableau 14 (Clotault, Thuillet *et al.*, 2012).

5.1.7 Identification des outliers

Des contraintes temporelles et techniques nous ont conduit à utiliser deux stratégies pour identifier les outliers. La distribution du Kst sous le modèle d'évolution défini par l'approche ABC a été simulée, et cette distribution a été comparée aux valeurs de Kst obtenues. Pour les D de Tajima calculés au sein des sauvages et des cultivés et le rapport des paramètres mutationnels cultivés / sauvages (theta), les distributions simulées sous le modèle dévolution défini par l'ABC n'ont pas pu être calculées pour des raisons techniques, une approche exploratoire a été adoptée.

Dans le tableau 15 figure le nombre d'outliers en fonction des différents intervalles de confiance pour le Kst. On constate que la distribution du Kst n'est pas centrée. Il est fort probable que ceci soit du à une mauvaise estimation du Kst. En effet, les outliers sont beaucoup plus nombreux pour des valeurs proches de zéro. Les outliers détectés avec les différentes statistiques se trouvent en annexes 15, 16, 17 et 18.

Tableau 14 : Récapitulatif de la variation des paramètres estimés en fonction du taux de mutation

Paramètre	μ = 1.0e-08	μ = 3.3e-08	μ = 7.90e-09
Taille de la population sauvage (N_0)	26 250	7 955	33 228
Taille de la population cultivée (N_1)	16 958	5 139	21 465
Date du goulot d'étranglement (années/$4N_0$)	2 946	890	3 721
Durée du goulot d'étranglement (année/$4N_0$)	2 735	827	3 455
Force du goulot d'étranglement	16 806	5 093	21 274
Flux migratoire bidirectionnel	0,00000507	0,00001673	0,000004

μ = 1.0e-08 (Hamblin et al., 2006) ; μ = 3.3e-08 et μ = 7.90e-09 (Clotault et al., 2012)

Tableau 15 : Nombre d'outliers sélectionné en fonction des intervalles de confiance retenus

	Nombre
Gènes avec valeur de Kst	14150
Gènes avec valeur de Kst faibles > 95%	1276
Gènes avec valeur de Kst faibles > 99%	117
Gènes avec valeur de Kst faibles > 99.9%	42
Gènes avec valeur de Kst élevées > 95%	22
Gènes avec valeur de Kst élevées > 99%	2
Gènes avec valeur de Kst élevées > 99.9%	0

6 Discussion

6.1 Identification et sélection des polymorphismes utilisés pour les analyses évolutives

6.1.1 Un échantillonnage pertinent et améliorable

Cette étude porte sur un échantillon de 21 génotypes. Les critères de sélection privilégiés pour le choix des accessions s'appuient sur la couverture des différentes races et la diversification des origines. Les génotypes couvrent la diversité des deux compartiments recensée actuellement de manière satisfaisante comme les analyses de structure exposées dans l'encadré 1 le prouvent. Cependant, l'analyse est restreinte à un noyau d'effectif faible non adapté pour estimer convenablement la structure qui existe au sein des deux groupes. Pour cette raison, il sera impossible d'explorer la pertinence de l'existence d'évènements de domestication multiples.

De plus, parmi les accessions sauvages, une se révèle plus proche des accessions cultivées (IS14719) (encadré 1, arbre F), ce positionnement atypique ayant été confirmé avec les polymorphismes identifiés dans cette étude. Un semi de ce génotype en serre a révèlé un phénotype cultivé et un poids de grains nettement supérieur à la moyenne du poids des grains des individus sauvages (fig. 9). Une analyse de structure (fig. 10) avec les marqueurs identifiés dans notre étude révèle la possibilité d'une confusion dans les semences d'IS14719 qui est proche d'un individu de la race *kafir* (IS29407) (fig. 11) originaire d'Afrique du Sud alors qu'il est censé venir d'Ethiopie (annexe 19). On note que cette analyse permet également d'identifier un autre génotype (SSM1057) cultivé apparenté au groupe des sauvages. Cet individu appartient à la race *guinea marga* qui est considérée comme intermédiaire entre sauvages et cultivés.

6.1.2 Une couverture du transcriptome satisfaisante et optimisable

L'extraction de la partie exprimée du génome à partir du grain, des feuilles et fleurs a été entreprise dans cette étude car on considère qu'elle est à l'origine de la majorité des facteurs influençant la variabilité des caractères phénotypiques (Sato, Antonio *et al.*, 2011; Schmid, Davinson *et al.*, 2011). Pour le séquençage, le budget relativement réduit, a également motivé cette décision (1 run illumina pour 10 accessions).

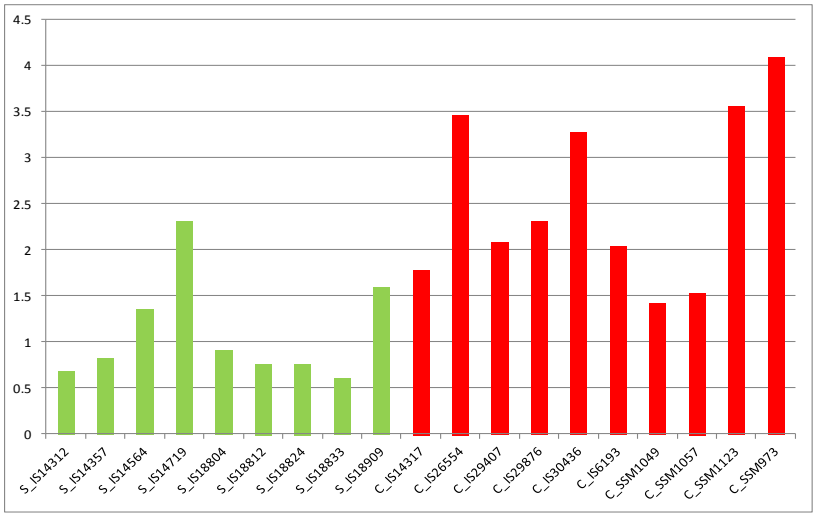


Figure 9 : Poids de 100 grains (en grammes) des génotypes sauvages en vert et cultivés en rouge. Un individu sauvage absent dont on ne disposait pas des graines pour en évaluer la masse.

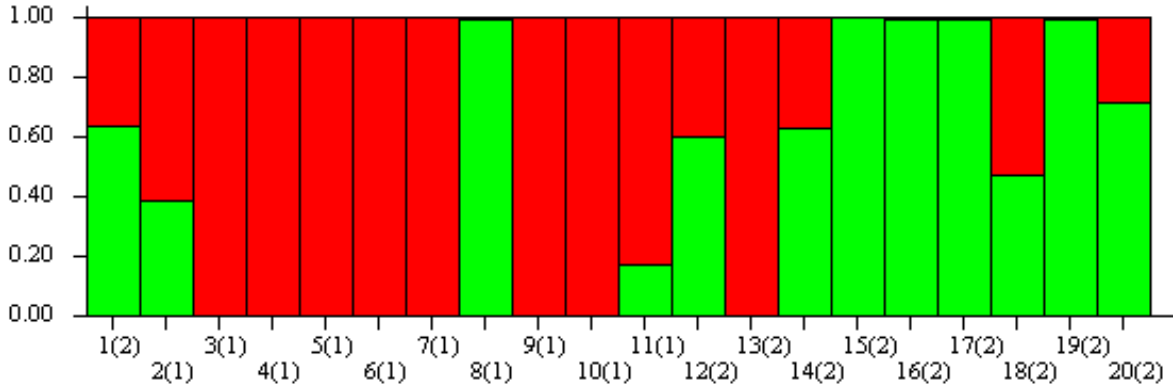


Figure 10 : Résultats de l'analyse de structure réalisée avec 16 122 SNP choisi aléatoirement pour les 10 individus cultivés à gauche et les 10 individus sauvages à droite. En vert, génotype sauvage et en rouge génotype cutlivé (K=2).

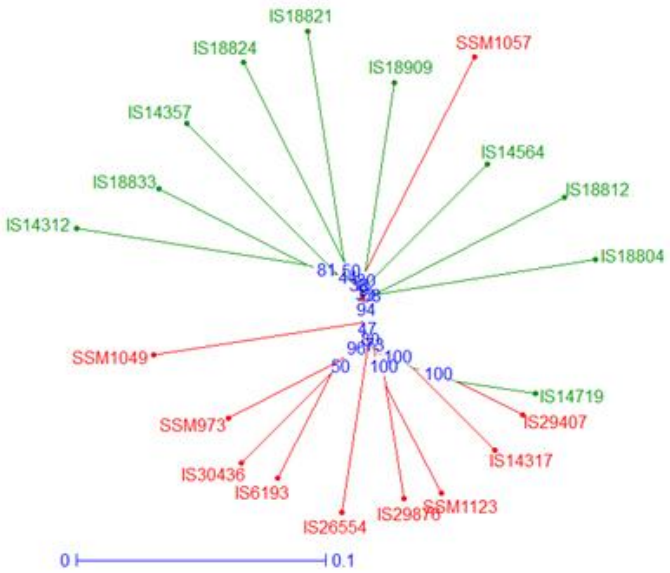


Figure 11 : Arbre phylogénétique illustrant la structure existante entre les 20 individus de notre étude

Ce choix paraît pertinent pour les raisons suivantes. D'une part, parce que la couverture des 27 609 gènes présents dans l'annotation est relativement complète (seulement 917 gènes n'accueillent aucun read). D'autre part, hormis les contaminants, on suppose que des gènes sont non annotés (40% des séquences non mappées ne sont pas annotées).

Cependant, malgré cette couverture étendue, on constate que le jeu de polymorphismes fiables (8X de profondeur minimum) se limite à 16 122 contigs soit 58% des gènes au total, sans compter les gènes non annotés. Ce défaut de couverture pourrait altérer la représentativité du jeu final sachant que le séquençage n'est pas équitable pour les deux compartiments. En effet, suite au premier nettoyage on dispose de près de 2 fois plus de séquences sauvages que de cultivées dont la répartition par individus après mapping (fig. 12) est plus homogène que dans le compartiment sauvage. L'outgroup présente 5 fois plus de séquences que certains individus d'où le fait qu'une grande partie des polymorphismes lui soit propre. Ces déséquilibres peuvent avoir des conséquences sur les résultats finaux qu'il serait pertinent d'évaluer en construisant, par exemple, un nouveau modèle d'évolution du sorgho à partir d'un jeu de séquences pour lesquelles on dispose des séquences pour l'ensemble des individus (soit des alignements de 40 séquences). Pour obtenir un jeu de séquences plus représentatif du transcriptome et dans un contexte de changement climatique il serait intéressant de mener une nouvelle étude en intégrant le système racinaire dans les tissus prélevés à différents stades phénologiques.

Plus de 2 millions de séquences supplémentaires sont mappées sur le génome par rapport au transcriptome. Soit 26% des SNP PASS en zone non annotée (hors mRNA). Dans une fenêtre de 500 pb autour des gènes on retrouve 30% de ces polymorphismes. On confirme donc la révision nécessaire de l'annotation du génome. Dans un premier temps il paraissait plus pertinent de constituer le set de SNP final à partir de ceux issus du mapping sur la référence génome sachant aussi que 18% des SNP PASS se trouvent en zone intronique. Contrairement au programme TopHat (Trapnell, 2012) qui sectionne facilement les reads lorsque les mismatch* sont trop nombreux, BWA utilisé dans cette étude a tendance à les conserver et génère une fréquence de SNP clusters faux positifs aux frontières intron/exons très élevée. Parmi ces derniers peuvent se trouver des polymorphismes vrais qui seront éliminés involontairement du set définitif. La solution alternative aurait consisté à combiner les SNP introniques issus du mapping sur génome avec les SNP exoniques issus du transcriptome. Vu les difficultés de conversion des coordonnées et la présence des SNPClusters on ne conservera que les SNP issus du mapping sur transcriptome.

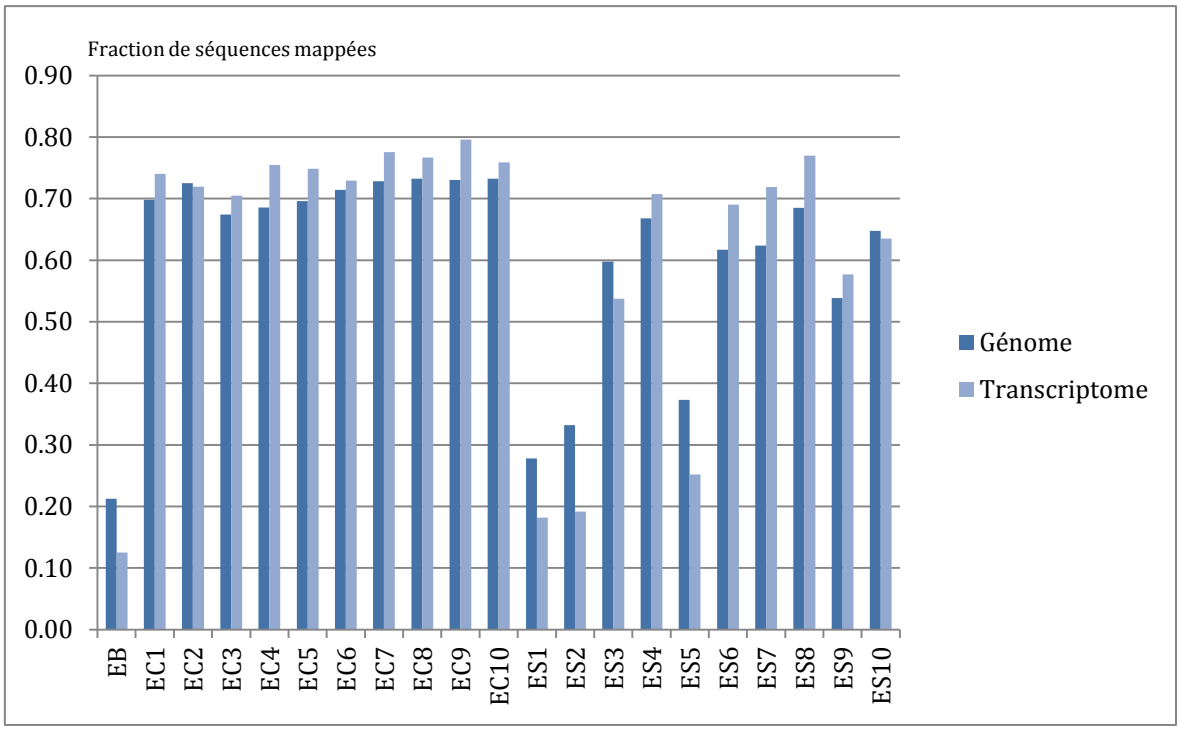


Figure 12 : Fraction des séquences mappées par individus

L'évaluation des taux de faux positifs et faux négatifs n'étant qu'approximative, nous avons envisagé d'appliquer un certain nombre de filtres pour la mise au point du jeu définitif. Il est possible d'éliminer une part importante des faux positifs avec la recalibration. Le calcul de l'algorithme se fait à partir de critères modulables et il est possible de faire varier la confiance que l'on accorde aux différents sets de SNP source. On constate qu'un Ti/Tv faible (1,01) est associé à un jeu de polymorphismes de moins bonne qualité et qu'un Ti/Tv élevé est caractéristique d'un set de SNP plus fiables (1,70 et 1,53 pour les SNP PASS du transcriptome et génome). Cependant, même si le programme GATK l'utilise, il n'est pas suffisant à lui seul pour estimer le taux de faux positifs précisément. Dans notre cas, les origines des SNP sources (rapport entre les SNP issus du compartiment sauvage et ceux issus du compartiment cultivé) n'étaient en plus pas représentatives du jeu obtenu lors de la présente étude. Le faible nombre de SNP sources, majoritairement issus d'individus cultivés, utilisés dans cette démarche ne permet pas de minimiser une perte non négligeable de SNP PASS valides, appartenant au compartiment sauvage et spécifique de l'outgroup, réduisant ainsi le set définitif à un noyau restreint et déséquilibré. L'incompatibilité des jeux de SNP sources et à recalibrer nous a définitivement fait renoncer à cette option de filtration.

En ce qui concerne les polymorphismes induits par le mapping de reads issus de paralogues sur une même référence, nous avons opté pour une approche basée sur l'élimination des polymorphismes présentant des taux d'individus hétérozygotes trop important (> 50 % pour les individus cultivés et > 70 % pour les individus sauvages). Il faut néanmoins noter, que les autres polymorphismes présents dans ces gènes qui ne présentaient pas ces biais ont été conservés (178 337 polymorphismes issus de 4332 gènes). Cette conservation est critiquable dans un contexte d'analyse évolutive direct sur les informations de séquences obtenues. En effet dans ce cas, une augmentation artificielle du niveau d'hétérozygotie est engendrée induisant des biais dans les analyses évolutives. Il pourrait par contre être pertinent de les tester dans des études d'association.

En contrepartie, l'élimination des polymorphismes fortement hétérozygotes du set définitif induit potentiellement l'exclusion de gènes affectés par de la sélection balancée qui sont potentiellement intéressants.

6.2 La définition du modèle d'évolution du sorgho

Le modèle DOM correspond à un scénario de domestication simple. Le flux de gènes entre les compartiments n'est pas modulable au cours du temps alors qu'on pense qu'il était d'abord orienté du compartiment sauvage vers le compartiment cultivé dans un premier temps, et il serait toujours asymétrique mais probablement inversé de nos jours (Barnaud, Deu *et al.*, 2009; Mutegi, Sagnard *et al.*, 2012). Les paramètres obtenus à partir des priors de loi uniforme (probabilité constante d'apparition de tous les intervalles du support) concordent bien avec la littérature (Hamblin, Casa *et al.*, 2006). Notamment pour ce qui est de la date et la durée du goulot d'étranglement. La population issue du goulot d'étranglement a bien subie une expansion, que l'on estime relativement faible, gain de moins de 1% de la population ayant subie le goulot d'étranglement. La contribution à cette expansion est certainement plus originaire de la dispersion du sorgho dans le monde que du flux de gène extrêmement faible entre les deux compartiments. Le faible nombre de migrants, soit 1 migrant toutes les deux générations, est en accord avec les études précédentes réalisées sur le haricot (Mamidi, Rossi *et al.*, 2011) ou les espèces cultivés (Papa, 2005). C'est la variation, même très faible, du taux de mutation qui pourrait induire un biais et qu'il aurait été intéressant d'estimer dans notre étude.

6.3 Identification des gènes impliqués dans la domestication ou d'intérêt adaptatifs

6.3.1 La stratégie de détection des outliers est perfectible

Les SNP de faible fréquence sont plus facilement identifiables sur des larges panels d'individus. Notre étude est réalisée sur un panel relativement restreint, ce qui génère un excès d'allèles de haute fréquence (Ramirez-Soriano et Nielsen, 2009). Il pourrait être intéressant de suivre la démarche de Ramirez et al (2009) pour vérifier la pertinence de nos résultats qui consiste à mettre en œuvre des méthodes plus robustes basées sur la structure haplotypique ou la correction des estimateurs statistiques. Il est également possible de simuler des données afin d'obtenir des distributions qui prennent en compte ce biais.

6.3.2 Une détection pertinente à partir d'une distribution soumise au modèle neutre

La première stratégie, permet d'identifier des outliers en fonction du Kst et s'appuie sur sa distribution révisée par le modèle DOM. Des intervalles de confiance se dégagent deux profils de gènes. Les gènes à Kst élevé qui signifie que la différenciation entre les compartiments est importante et les gènes caractérisés par un Kst faible qui présentent des profils de diversité comparables entre les deux compartiments. Le nombre élevé de ces derniers (annexe 16) est probablement lié aux difficultés d'estimation du Kst pour les faibles valeurs. Parmi les outliers appartenant à cette catégorie, on remarque la présence d'un gène restaurateur de fertilité (Sb02g004530). Parmi les outliers présentant un Kst élevé, on trouve des gènes intervenant dans le contrôle de la croissance cellulaire (Sb06g021260), de la réponse aux nutriments et facteurs de croissance (Sb01g008695) et dans la réponse au stress. Notamment un homologue de TOR qui est à l'origine d'une production accrue de biomasse chez Arabidopis. Ce gène pourrait donc être un candidat potentiel à l'origine de la différence de production de biomasse qui existe entre les individus sauvages et cultivés. Dans cette même catégorie, on retrouve également des gènes impliqués dans la synthèse de la paroi cellulaire (Sb06g021260 ; Sb10g001710[3]) probablement responsable de différences importantes dans la qualité de la biomasse entre les deux compartiments. Des gènes de résistance aux maladies sont aussi mis en évidence. Par exemple, GH3-8, un gène de réponse à l'auxine active la résistance par l'activation du signal de l'acide salicylique (Ding, Cao *et al.*, 2008). Il est intéressant de noter que Sb01g019560 outlier de la même catégorie a aussi été identifié comme outlier par Bouchet et al (2012) sur la base de sa différenciation entre les différents groupes génétiques de sorgho cultivés. Suite à la recherche d'une homologie, on se rend compte qu'une calmodulin binding protein est prédite pour ce gène.

6.3.3 Une seconde stratégie exploratoire

La seconde stratégie qui vise à identifier des outliers via les valeurs extrêmes de D de Tajima a surtout été réalisée dans un but exploratoire afin de donner une idée de la fonction des gènes mis en évidence. En effet, la distribution de l'estimateur n'est pas soumise au modèle neutre et seul un nombre de valeurs extrêmes déterminé au préalable est retenu. Les statistiques simulées via Egglib ont en effet été calculées sur l'effectif global et non sur les compartiments de manière distincte.

[3] Pectine esterase activity

La recherche des outliers a été réalisée sur les deux compartiments d'individus de façon indépendante. Pour les gènes sous sélection directionnelle (gènes d'adaptation) dans le compartiment cultivé (D<0) on peut porter notre attention sur Sb01g047830 correspondant probablement à une protéine modératrice d'une kinase impliquée dans diverses réponses à des stress environnementaux (Xue, Wang *et al.*). On note aussi que Sb02g027070 appartenant à cette même catégorie est impliqué dans le métabolisme de l'azote.

Dans le même compartiment, parmi les outliers ayant un D > 0 (sélection balancée), on retient Sb08g006680 impliqué dans la réponse à la sécheresse et au stress salin. On peut émettre l'hypothèse que ces résistances induisent certainement des coûts métaboliques élevés qui peuvent expliquer le fait que l'allèle ne soit pas fixé sans pression de sélection.

Concernant les outliers du compartiment sauvage, dans la première catégorie, on retient Sb01g048100 impliqué dans des processus de défense (Kaku, Nishizawa *et al.*, 2006) et d'autres gènes impliqués dans la synthèse des parois.

En complément de l'analyse des pressions évolutive intra compartiment, et de l'analyse de leur différenciation on se propose d'identifier des outliers à partir du rapport θ_c/θ_s. Il s'agit ici surtout de retrouver les gènes présentant un rapport θ_c/θ_s faible témoins d'une baisse de la diversité suite à la domestication. Les gènes qui présentent un rapport θ_c/θ_s élevé sont probablement issus de phénomènes d'adaptation ou de la sélection artificielle et ont « acquis » des allèles supplémentaires en réponse à la sélection artificielle ou afin de s'adapter à de nouvelles conditions pédo climatique. Sur la base de cette différenciation, on parvient à extraire des gènes impliqués dans les réponses hormonales et à la lumière. Mais aussi des gènes qui jouent un rôle dans la synthèse de molécules de stockage et de structure sur lesquelles on peut porter une attention particulière pour les programmes destinés à l'amélioration de la biomasse.

Par exemple, dans la catégorie d'outliers qui présentent un ratio faible on identifie des gènes de réponse à l'auxine (Sb03g034850 et Sb06g011767) qui sont impliqués dans l'organogénèse et donc l'architecture des plantes. Ces gènes sont probablement impliqués dans la perte du tallage ou les modification de l'architecture de la plante. On note qu'une signature de sélection pour une ARF* a été mise en évidence chez *Pinus taeda* (Ersoz, Wright *et al.*, 2010). Dans cette même catégorie, d'autre gènes impliqués dans la réponse au stress et la synthèse des parois cellulaires ont également été detectés (Sb01g041880).

Un gène potentiellement impliqué dans la morphogénèse et le développement nommé Homeobox-leucine zipper protein ROC3 (Sb01g028160) est identifié dans les outliers qui présentent un ratio élevé. On retrouve également des gènes ayant un rôle dans la biosynthèse de la cellulose des parois (en relation avec les stress biotiques et abiotiques).

Afin de consolider cette analyse il serait nécessaire de procéder à la comparaison des distributions de l'estimateur D de Tajima propres à chaque groupe et soumises au modèle de domestication.

Conclusion

A partir de 20 individus sauvages et cultivés nous avons pu identifier plus de 500 000 SNP et indels issus la portion exprimée du génome du sorgho. Grâce aux méthodes de coalescence, ce jeu de polymorphismes nous a permis de définir l'histoire évolutive du sorgho la plus probable. Sur cette base nous avons tenté de mettre en évidence des gènes présentant des patrons de diversité divergents des attendus neutres. L'analyse de l'indice de différenciation Kst entre les compartiments sauvage et cultivé nous a permis de détecter 64 gènes présentant des différenciations extrêmes par rapport aux attendus neutres. Ces gènes interviennent notamment dans le contrôle de la croissance cellulaire, la réponse aux nutriments et aux facteurs de croissance ainsi que dans la réponse aux stress biotiques et abiotiques. Une analyse exploratoire basée sur les valeurs de D de Tajima à l'intérieur des compartiments cultivés et sauvages et sur les ratios de diversité entre ceux-ci ont permis de mettre en évidence d'autres gènes impliqués dans le développement des organes, la mise en place des parois cellulaires et la réponse aux stress. Les fonctions de ces gènes étant cohérentes avec la connaissance actuelle de la variabilité phénotypique entre les compartiments sauvages et cultivés. Une démarche de génétique quantitative et moléculaire sera maintenant nécessaire pour valider les effets de la variabilité de ces gènes sur des caractères d'intérêt pour les programmes de sélection actuels.

Comme nous l'avons abordé dans cette étude, il existe d'autres estimateurs basés sur la distribution allélique pour mettre en évidence les gènes de domestication ou d'intérêt adaptatif. Ainsi, le perfectionnement du modèle neutre, notamment par l'élimination des polymorphismes originaires de séquences paralogues, nous permettrait de confirmer les outliers déjà identifiés par notre analyse exploratoire.

Le phasage de notre jeu de polymorphismes serait également pertinent pour comparer les outliers générés sur la base du Kst avec ceux identifiés via des tests basés sur la variabilité intraspécifique entre les différentes classes de mutations au sein d'un locus (mutation synonymes et non synonymes). Outre l'identification des gènes impliqués dans les processus de domestication ou d'adaptation, qu'il sera intéressant de réintégrer dans les nouveaux programmes de sélection le jeu de polymorphismes identifié dans cette étude a aussi un intérêt dans le cadre des approches de génétique quantitative en permettant une amélioration de la densité de marquage.

Bibliographie

Acot, O. J. B. (2002). The molecular study of the population genetic structure of hippopotamus (hippopotamus amphibius) in eastern and southern africa.p.

Barnaud, A.; M. Deu; E. Garine; J. Chantereau; J. Bolteu; E. O. Koida; D. McKey et H. I. Joly (2009). A weed-crop complex in sorghum: The dynamics of genetic diversity in a traditional farming system. *Am J Bot*, **96**, 1869-79.

Barro-Kondombo, C.; F. Sagnard; J. Chantereau; M. Deu; K. Vom Brocke; P. Durand; E. Goze et J. D. Zongo (2010). Genetic structure among sorghum landraces as revealed by morphological variation and microsatellite markers in three agroclimatic regions of Burkina Faso. *Theor Appl Genet*, **120**, 1511-23.

Beaumont, M. A.; W. Y. Zhang et D. J. Balding (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025-2035.

Bouchet, S.; D. Pot; M. Deu; J. F. Rami; C. Billot; X. Perrier; R. Rivallan; L. Gardes; L. Xia; P. Wenzl; A. Kilian et J. C. Glaszmann (2012). Genetic Structure, Linkage Disequilibrium and Signature of Selection in Sorghum: Lessons from Physically Anchored DArT Markers. *PLoS One*, **7**, e33470.

Brown, P.; S. Myles et S. Kresovich (2011). Genetic Support for Phenotype-based Racial Classification in Sorghum. *Crop Science*, **51**, 224-230.

Casa, A.; G. Pressoir; P. Brown; S. Mitchell; W. Rooney; M. Tuinstra; C. Franks et S. Kresovich (2008). Community Resources and Strategies for Association Mapping in Sorghum. *Crop Sci*, **48**, 30-40.

Casa, A. M.; S. E. Mitchell; M. T. Hamblin; H. Sun; J. E. Bowers; A. H. Paterson; C. F. Aquadro et S. Kresovich (2005). Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. *Theor Appl Genet*, **111**, 23-30.

Clotault, J.; A. C. Thuillet; M. Buiron; S. De Mita; M. Couderc; B. I. Haussmann; C. Mariac et Y. Vigouroux (2012). Evolutionary history of pearl millet (Pennisetum glaucum [L.] R. Br.) and selection on flowering genes since its domestication. *Mol Biol Evol*, **29**, 1199-212.

David, J.; O. Loudet et J.-C. Glaszmann (2006). Le regard de la génomique sur la diversité naturelle des plantes cultivées. *Biofutur*, **266**, 22-27.

de Alencar Figueiredo, L. F.; C. Calatayud; C. Dupuits; C. Billot; J. F. Rami; D. Brunel; X. Perrier; B. Courtois; M. Deu et J. C. Glaszmann (2008). Phylogeographic evidence of crop neodiversity in sorghum. *Genetics*, **179**, 997-1008.

De Mita, S. et M. Siol (2012). EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genetics*, 13:27.

de Wet, J. M. J. et J. R. Harlan (1971). The origin and domestication of Sorghum bicolor. *Economic Botany*, **25**, 128-135.

Deu, M.; H. F. Rattunde et J. Chantereau (2006). A global view of genetic diversity in cultivated sorghums using a core collection. *Genome*, **49**, 168-180.

Deu, M.; F. Sagnard; J. Chantereau; C. Calatayud; D. Hérault; C. Mariac; J. L. Pham; Y. Vigouroux; I. Kapran; P. Traore; A. Mamadou; B. Gerard; J. Ndjeunga et G. Bezançon (2008). Niger-wide assessment of in situ sorghum genetic diversity with microsatellite markers. *Theor Appl Genet*, **116**, 903-913.

Ding, X.; Y. Cao; L. Huang; J. Zhao; C. Xu; X. Li et S. Wang (2008). Activation of the Indole-3-Acetic Acid–Amido Synthetase GH3-8 Suppresses Expansin Expression and Promotes Salicylate- and Jasmonate-Independent Basal Immunity in Rice. *The Plant cell*, **20**, 228–240.

Dje, Y.; M. Heuertz; C. Lefebvre et X. Vekemans (2000). Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. *Theoretical and Applied Genetics*, **100**, 918-925.

Doebley, J.; A. Stec et C. Gustus (1995). teosinte branched1 and the Origin of Maize: Evidence for Epistasis and the Evolution of Dominance. *Genetics Society of America*, **141**, 333-346.

Doebley, J.; A. Stec et L. Hubbard (1997). The evolution of apical dominance in maize. *Nature*, **386**, 485-488.

Doggett, H. (1988). Sorghum, Longman Scientific and Technical., London. Book-

Ersoz, E. S.; M. H. Wright; S. C. Gonzalez-Martinez; C. H. Langley; D. B. et A. D. Neale (2010). Evolution of Disease Response Genes in Loblolly Pine: Insights from Candidate Genes. *PLoS One*, **5**.

Excoffier, L.; G. Laval et S. Schneider (2005). Arlequin version 3.0: an integrated software package for population genetics data analysis. *Evol Bioinf Online*, **1**, 47-50.

Frere, C. H.; P. J. Prentis; E. K. Gilding; A. M. Mudge; A. Cruickshank et I. D. Godwin (2011). Lack of Low Frequency Variants Masks Patterns of Non-Neutral Evolution following Domestication. *Plos One*, **6**.

Glemin, S. Les consequences evolutives des systemes de reproduction et d'autres traits d'histoire de vie. Montpellier.

Grenier, C.; P. Hamon et P. J. Bramel-Cox (2001). Core Collection of Sorghum: II. Comparison of Three Random Sampling Strategies. *Crop Sci*, **41**, 241-246.

Gurian-Sherman, D. (2009). Failure to yield, Evaluating the Performance of Genetically Engineered Crops. *Union of Concerned Scientists*, 1-51.

Hamblin, M. T.; A. M. Casa; H. Sun; S. C. Murray; A. H. Paterson; C. F. Aquadro et S. Kresovich (2006). Challenges of detecting directional selection after a bottleneck: lessons from Sorghum bicolor. *Genetics*, **173**, 953-64.

Hamblin, M. T.; S. E. Mitchell; G. M. White; J. Gallego; R. Kukatla; R. A. Wing; A. H. Paterson et S. Kresovich (2004). Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics*, **167**, 471-83.

Harlan, J. R. (1995). The Living Fields, Our Agricultural Heritage. Cambridge, UK.

Hudson, R. R.; M. Slatkint et W. P. Maddison (1992). Estimation of Levels of Gene Flow From DNA Sequence Data. *Genetics Society of America*, **132**, 583-589.

Kaku, H.; Y. Nishizawa; N. Ishii-Minami; C. Akimoto-Tomiyama; N. Dohmae; K. Takio; E. Minami et N. Shibuya (2006). Plant cells recognize chitin fragments for defense signaling through a plasma membrane receptor. *PNAS*, **103**, 11086-11092.

Mamidi, S.; M. Rossi; D. Annam; S. Moghaddam; R. Lee; R. Papa et P. McClean (2011). Investigation of the domestication of common bean (Phaseolus vulgaris) using multilocus sequence data. *Functional Plant Biology*, **38**, 953–967.

McKenna, A.; M. Hanna; E. Banks; A. Sivachenko; K. Cibulskis; A. Kernytsky; K. Garimella; D. Altshuler; S. Gabriel; M. Daly et M. DePristo (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**, 1297-1303.

Muraya, M. M.; S. de Villiers; H. K. Parzies; E. Mutegi; F. Sagnard; B. M. Kanyenji; D. Kiambi et H. H. Geiger (2011). Genetic structure and diversity of wild sorghum populations (Sorghum spp.) from different eco-geographical regions of Kenya. *Theor Appl Genet*, **123**, 571-83.

Muraya, M. M.; E. Mutegi; H. H. Geiger; S. M. de Villiers; F. Sagnard; B. M. Kanyenji; D. Kiambi et H. K. Parzies (2011). Wild sorghum from different eco-geographic regions of Kenya display a mixed mating system. *Theor Appl Genet*, **122**, 1631-9.

Mutegi, E.; F. Sagnard; M. Labuschagne; L. Herselman; K. Semagn; M. D. S. d. Villiers; B. M. Kanyenji; C. N. Mwongera; P. C. S. Traore et D. Kiambi (2012). Local scale patterns of gene flow and genetic diversity in a crop–wild–weedy complex of sorghum (Sorghum bicolor (L.) Moench) under traditional agricultural field conditions in Kenya. *Conserv Genet*.

Mutegi, E.; F. Sagnard; K. Semagn; M. Deu; M. Muraya; B. Kanyenji; S. de Villiers; D. Kiambi; L. Herselman et M. Labuschagne (2011). Genetic structure and relationships within and between cultivated and wild sorghum (Sorghum bicolor (L.) Moench) in Kenya as revealed by microsatellite markers. *Theor Appl Genet*, **122**, 989-1004.

Papa, R. (2005). Gene flow and introgression between domesticated crops and their wild relatives. *THE ROLE OF BIOTECHNOLOGY*.

Paterson, A. H.; J. E. Bowers; R. Bruggmann; I. Dubchak; J. Grimwood; H. Gundlach; G. Haberer; U. Hellsten; T. Mitros; A. Poliakov; J. Schmutz; M. Spannagl; H. Tang; X. Wang; T. Wicker; A. K. Bharti; J. Chapman; F. A. Feltus; U. Gowik; I. V. Grigoriev; E. Lyons; C. A. Maher; M. Martis; A. Narechania; R. P. Otillar; B. W. Penning; A. A. Salamov; Y. Wang; L. Zhang; N. C. Carpita; M. Freeling; A. R. Gingle; C. T. Hash; B. Keller; P. Klein; S. Kresovich; M. C. McCann; R. Ming; D. G. Peterson; R. Mehboob ur; D. Ware; P. Westhoff; K. F. Mayer; J. Messing et D. S. Rokhsar (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551-6.

Pritchard, J. K. et N. A. Rosenberg (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, **65**, 220-8.

Pulchérie Barro-Kondombo, C.; K. Vom Brocke; J. Chantereau; F. Sagnard et J. D. Zongo (2008). Phenotypic variability of local sorghum cultivars/varieties of two regions in Burkina Faso: The Boucle du Mouhoun and the Centre West.

Purugganan, M. et D. Fuller (2009). The nature of selection during plant domestication. *Nature*, **12**, 843-848.

Ramirez-Soriano, A. et R. Nielsen (2009). Correcting Estimators of u and Tajima's D for Ascertainment Biases Caused by the Single-Nucleotide Polymorphism Discovery Process. *Genetics Society of America*, **181**, 701–710.

Ross-Ibarra, J.; P. L. Morrell et B. S. Gaut (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc Natl Acad Sci U S A*, **104 Suppl 1**, 8641-8.

Rozas, J.; J. C. Sanchez-DelBarrio; X. Messeguer et R. Rozas (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496-2497.

Sagnard, F.; M. Deu; D. Dembele; R. Leblois; L. Toure; M. Diakite; C. Calatayud; M. Vaksmann; S. Bouchet; Y. Malle; S. Togola et P. C. Traore (2011). Genetic diversity, structure, gene flow and evolutionary relationships within the Sorghum bicolor wild-weedy-crop complex in a western African region. *Theor Appl Genet*, **123**, 1231-46.

Sahoo, L.; J. J. Schmidt; J. F. Pedersen; D. J. Lee et J. L. Lindquist (2010). Growth and fitness components of wild × cultivated sorghum bicolor (poaceae) hybrids in nebraska. *American Journal of Botany*, **97**, 1610–1617.

Sato, Y.; B. Antonio; N. Namiki; R. Motoyama; K. Sugimoto; H. Takehisa; H. Minami; K. Kamatsuki; M. Kusaba; H. Hirochika et Y. Nagamura (2011). Field transcriptome revealed critical developmental and physiological transitions involved in the expression of growth potential in japonica rice. *BMC Plant Biology*, **11**, 1-15.

Schmid, M.; T. Davinson et S. Henz (2011). A gene expression map of Arabidopsis thaliana development. *Nature genetics*, **37**, 501-509.

Tajima, F. (1989). Statistical Methods for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, **123**, 585-595.

Tesso, T.; I. Kapran; C. Grenier; A. Snow; P. Sweeney; J. Pedersen; D. Marx; G. Bothma et G. Ejeta (2008). The potential for crop-to-wild gene flow in sorghum in Ethiopia and Niger: A geographic survey. *Crop Science*, **48**, 1425-1431.

Trouche, G. et J. Chantereau (2009). Problématiques de sélection du sorgho comme culture multi-usage.

Wang, J.; B. Roe; S. Macmil; Q. Yu; J. E. Murray; H. Tang; C. Chen; F. Najar; G. Wiley; J. Bowers; M. A. Van Sluys; D. S. Rokhsar; M. E. Hudson; S. P. Moose; A. H. Paterson et R. Ming (2010). Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics*, **11**, 261.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Xue, T.; D. Wang; S. Zhang; J. Ehlting; F. Ni; S. Jakab; C. Zheng et Y. Zhong Genome-wide and expression analysis of protein phosphatase 2C in rice and Arabidopsis. *BMC Genomics*, **9**, 1-21.

Zheng, L. Y.; X. S. Guo; B. He; L. J. Sun; Y. Peng; S. S. Dong; T. F. Liu; S. Jiang; S. Ramachandran; C. M. Liu et H. C. Jing (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biol*, **12**, R114.

Sitographie

Babraham (page consultée le 7/07/2012) URL : http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

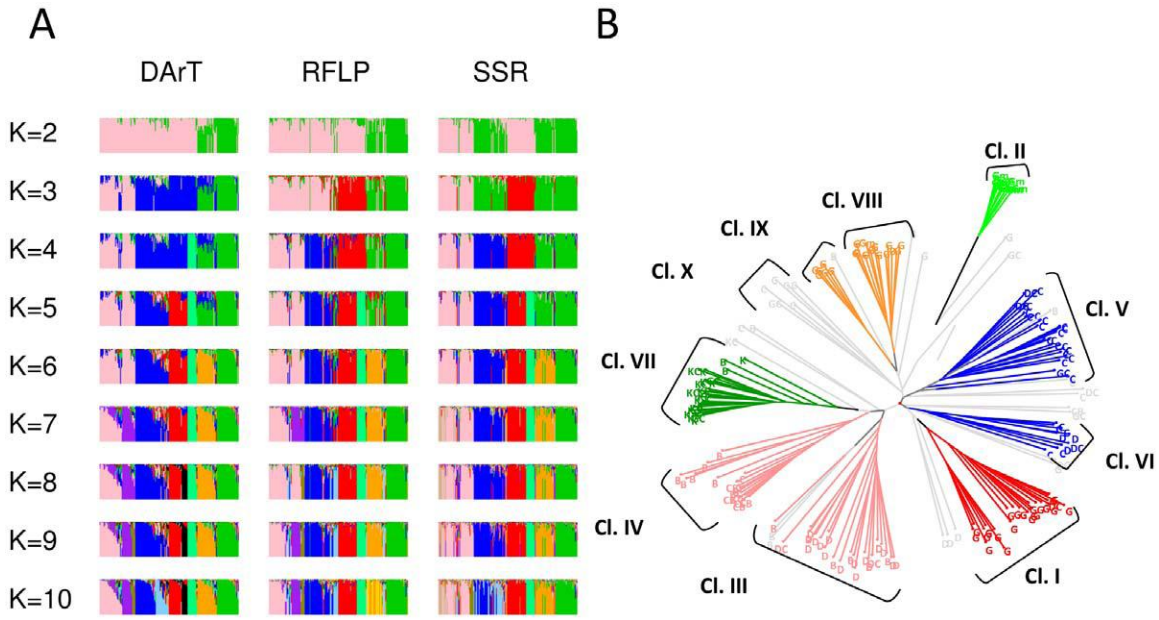
Carr (page consultée le 7/07/2012) URL : http://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html

FAO (page consultée le 7/07/2012) URL : http://www.fao.org/docrep/004/y2775f/y2775f09.htm

lh3 (page consultée le 7/07/2012) URL : http://seqanswers.com/forums/showthread.php?t=6854

Annexes

Annexe 1 : Structure génétique des sorghos cultivés à l'échelle mondiale



Identification de groupes génétiques basée sur une analyse de structure grâce à différents types de marqueurs et comparaison du modèle le plus pertinent (K =6) avec une analyse de distances. (A) Composition du génome des accessions (en abscisse) en fonction de K qui représente le nombre de groupes hypothétiques qui composent la collection. Différents jeux de données ont été testés : 713 DArTs, 60 RFLPs, and 40 SSRs. (B) Arbre représentant les différents groupes dégagés de l'échantillon grâce aux similarités génétiques. Ces analyses sont basées sur un panel représentatif des sorghos cultivés mis au point à l'aide de critères de race, d'origine, de réponse à la photopériode et de méthode de culture.

Le groupe A comprend en rose les *durra* et *bicolor* de l'Inde, les *caudatum* et *caudatum bicolor* de Chine. Le groupe B en bleu comprend les *caudatum* et *durra* d'Afrique. Le groupe C en rouge comprend les *guinea* de l'Afrique occidentale. Le groupe D en vert clair comprend les *guinea margaritiferum* de l'Afrique de l'Ouest. Le groupe E en orange comprend les *guinea* de l'Afrique australe et de l'Asie. Le groupe F en vert comprend *kafir* et *kafir caudatum* de l'Afrique Australe. Les clusters* ont été identifiés par Deu et al. (2006).

Annexe 2 : Protocole d'échantillonnage

Les 20 individus étudiés ont été sélectionnés afin de maximiser la couverture de la diversité (en terme d'allèles capturés et de nombre d'haplotypes). Pour cela, les génotypes sont comparés deux à deux et seuls les plus éloignés sont conservés dans le set définitif via le logiciel DARwin.

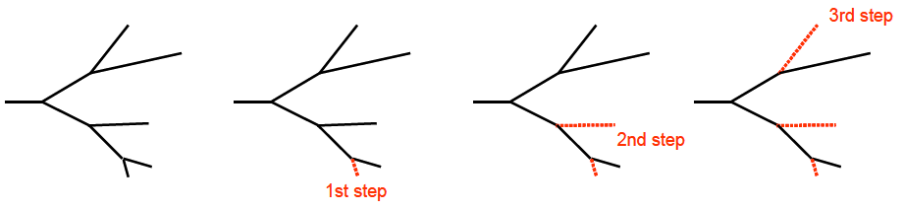


Schéma de sélection des génotypes pour maximiser la couverture de la diversité via le logiciel DARwin. Les 3 étapes de gauche à droite montrent la méthodologie de sélection des génotypes les plus éloignés les uns des autres.

Il est possible de poser des conditions et de prioriser la conservation de certains génotypes jugés par exemples pertinents pour la maximisation de la diversité. C'est le cas pour 10 d'entre eux sélectionnés sur un critère principal de profondeur et de qualité de séquençage.

Annexe 3 : Protocole d'extraction de l'ARN de feuilles

Extraction d'ARN Feuilles_Tiges_Racines

- Mettre environ 100 mg de matériel broyé à l'azote liquide dans un tube 2 ml
- Ajouter 1 ml de TRIzol pour 100 mg de tissus
- Homogénéiser
- Laisser 10 min à RT sous agitation
- Ajouter 0,2 ml de chloroforme par ml de TRIzol
- Incuber 10 min sous agitation
- Centrifuger 15 min (12000g, 4°C)
- Transférer le surnageant dans un nouveau tube
- Ajouter 0.5 ml d'isopropanol par ml de TRIzol
- Incuber 15 min à –20°C
- Centrifuger 10 min (12000g, 4°C)
- Prélever le surnageant et ajouter 1 ml d'éthanol 75% par ml de TRIzol
- Vortexer et centrifuger 5min (7500g, 4°C)

Sécher les culots (possibilité de faire le vide mais pas de rotor au speed vac) et reprendre dans 40µl d'eau Rnase free.

Annexe 4 : Protocole d'extraction de l'ARN de graines

Extraction d'ARN de graines

Etape 1

- Environ 100-150 mg de graines broyées
- Rajouter 500 µL de tampon d'extraction et agiter fortement
- Ajouter 335 µL de phénol saturé d'eau (agitation forte et incubation pendant 10 min à température ambiante)
- Ajouter 165 µL de chloroforme (agitation forte et incubation pendant 5 min)
- centrifugation (15000 x g, 20 min, température ambiante)
- la phase supérieure est transférée dans un nouveau tube

Etape 2

- ajout de 250 µL de phénol (agitation forte et incubation pendant 10 min à température ambiante)
- 250 µL de chloroforme (agitation forte et incubation pendant 5 min à température ambiante)
- centrifugation (15000 x g, 20 min, température ambiante)

Etape 3

- La phase supérieure issue de l'étape 2 est purifiée avec 500 µL de chloroforme
- Agitation forte et incuber 10 min à température ambiante
- centrifugation (15000 x g, 10 min, température ambiante)
- Les acides nucléiques contenus dans la phase aqueuse sont alors transférés dans de nouveaux tubes eppendorf
- précipités avec 1/10e V d'acétate de sodium 3 M pH 6.0 et 1 V d'isopropanol
- Le mélange est homogénéisé puis incubé 2 h à –20°C
- centrifugation (15000 x g, 20 min, 4°C),
- les culots d'acide nucléiques sont lavés avec 500 µL d'éthanol 70% (v/v)
- séchés et resuspendus dans 200 µL H2O
- Les ARN sont ensuite précipités spécifiquement au chlorure de lithium (2 M final) à -20°C, toute la nuit
- centrifugation (12000 x g, 30 min, 4°C)
- lavage des culots à l'éthanol 70% (v/v), les ARN sont repris dans 50 µL d'eau

Tampon d'extraction graines

(Tris-HCl 25 mM pH 8.0, NaCl 75 mM, SDS 1% (p/v) ; EDTA 25 mM pH 8.0)

Annexe 5 : Protocole d'extraction de l'ARN de fleurs

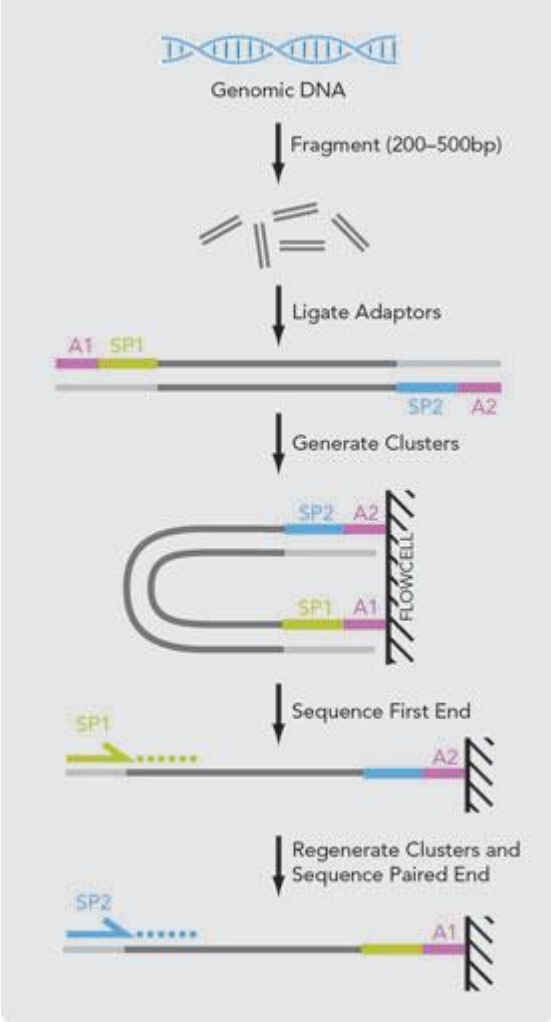
Extraction d'ARN de fleurs

- Environ 100 mg de fleurs broyées dans tube 2 ml.
- Rajouter 0.5 ml de tampon Plant RNA purification Reagent (Invitrogen)
- Homogénéiser
- Laisser 5 min à RT sous agitation
- Centrifuger 20 min (12000g, RT)
- Transférer le surnageant dans nouveau tube 1.5 ml
- Ajouter 0.1 ml de Na Cl 5M puis 0.3 ml de chloroforme
- Incuber 5 min sous agitation
- Centrifuger 10 min (12000g, 4°C)
- Transférer le surnageant dans un nouveau tube
- Ajouter 1V d'isopropanol
- Incuber 10 min à RT
- Centrifuger 10 min (12000g, 4°C)
- Prélever le surnageant et ajouter 1 ml d'éthanol 75%
- Vortexer et centrifuger 5min (7500g, 4°C)
- Sécher les culots et reprendre dans 40µl d'eau Rnase free.

Traitement à la désoxyribonucléase I (DNAse I promega)

- 40 µl d'ARN + 5µL de tampon 10X+ 4µl de DNAse I+ 1µl de RNAse inhibitor.
- Homogénéiser
- Incuber 30 min au bain marie à 37°C
- Transférer sur glace
- Ajuster le volume à 100 µl
- Purifier l'ARN avec le kit Nucleospin RNA plant (macherey-Nagel)
- Ajouter 300µl de RA1 + 300 µl d'éthanol 70%
- Mélanger et déposer sur colonne « cercle bleue »
- Centrifuger 1 min à 11000g
- Ajouter 350 µl de tampon MDB
- Centrifuger 1 min à 11000g
- Lavage de la membrane :
- Ajouter 200 µl de tampon RA2
- Centrifuger 30s à 11000g
- Ajouter 600 µl de tampon RA3
- Centrifuger 30s à 11000g
- Ajouter 250 µl de tampon RA3
- Centrifuger 2 min à 11000g
- Ajouter 50 µl d'eau RNase free
- Centrifuger 1 min à 11000g
- Dosage au nanoquant puis vérification de la qualité au bioanalyser.

Annexe 6 : Le séquençage paired-end



Le séquençage paired-end consiste à séquencer les reads à partir de leurs deux extrémités simultanément. On obtient ainsi les deux séquences appelées paired-end reads correspondant aux fragments 3' et 5' séquencés.

Annexe 7 : Récapitulatif des adaptateurs utilisés pour l'ensemble des 20 accessions

Adaptateurs	EC	ES	EB
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG	1		
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG		1	
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG	2	2	
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG	3	3	
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG	4	4	
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG	5	5	x
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG	6	6	
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTG	7	7	
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTG	8	8	
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG	9	9	
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG	10	10	

Annexe 8 : Détails des paramètres pour l'ensemble des programmes utilisés pour le nettoyage, mapping et détection des polymorphismes

Programme	Paramètres	Objectif
FastqC		Vérification de la qualité globale, peut être effectué après chaque étape de nettoyage
Cutadapt	Longueur sim >7 bases Longueur min = 20 Quality filter =25	Supprime les adaptateurs, similarité de 7 bases, longueur restante minimal de 20, toutes les bases restantes doivent avoir une qualité de 25 au minimum.
Filter	Longueur > 35 bases Qualité > 30	Elimine les reads trop courts (min. 35 b) et dont la qualité moyenne est trop faible (<30)
Compared Fastq paired		Compare les membres de deux fichiers sortis des filtres pour reconstituer les paires et filtrer les singles
Mapping	Mismatches = 5 Indels = 2	Mapping des reads sur une référence, avec une liberté de 5 erreurs et deux indels. (les contigs sont générés en toute fin de l'analyse en général) Single et paired end mappées séparément. Les deux fichiers de mapping par individus sont ensuite mergés (fusionnés).
Rmdup		Suppression des doublons (positions de mapping identiques, duplicats optiques ou PCR) – Doit être effectué à cette étape pour les paires. Est plus rapide ici que dans les fichiers initiaux (plus faciles de comparer 2x2 coordonnées que 76x76 bases)
Realigner Recalibrate		Realigner : réalignement des séquences autour des indels de bordure (anomalie induite par l'algorithme de mapping corrigée ici en SmithWaterman). Recalibrate : via des SNP validés, recalibrage de la qualité des polymorphismes mal considérés. Si un SNP détecté est déjà connu par une autre méthode/analyse, la position sera validée
Cleaner		Elimination des mapping ayant un score de MQ (mapping quality) trop faible, trop de mismatchs (3 max), indels (1 max) ou de positions possibles (une seule possible).
Merging bam		Fusion des fichiers individus
Genotyper call		Appel des variants SNP et indels par le GATK
Phasing		Ajout de l'information de phase si possible

Annexe 9 : Identification et définition des polymorphismes

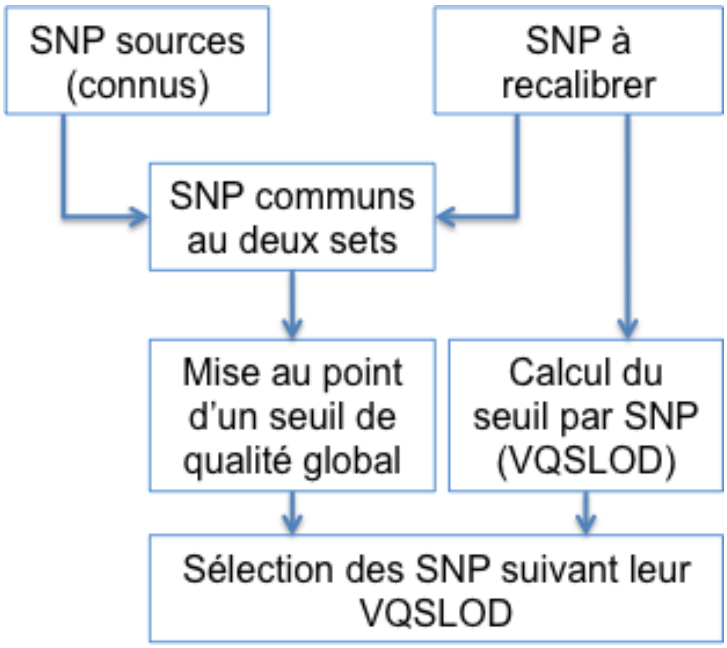
La détection des polymorphismes par Unified Genotyper se fait par une inférence en deux étapes. D'abord via une homologie entre les séquences individuelles et les fragments séquencés (reads assemblés) puis par la comparaison des fragments inférés à la référence de l'organisme. Ce système gère également la corrélation des erreurs entre les paires.

Assignation d'un filtre à chacun des polymorphismes par VariantFiltrationWalker

L'assignation des filtres par le VariantFiltrationWalker se fait sur la base de 5 informations propres aux polymorphismes de chacun des échantillons. Ces informations disponibles dans le FORMAT sont :

- Le génotype (GT). Pour un diploïde, le génotype indique les deux allèles portés par l'échantillon. Chacun d'eux est codé par 0 pour l'allèle de référence, 1 pour le premier alternatif, 2 pour la second alternatif… Quand il n'y a qu'un seul allèle alternatif (dans la très grande majorité des cas), le génotype peut être :
    - 0/0 : l'échantillon est homozygote de la référence ;
    - 0/1 : l'échantillon est hétérozygote, il porte une copie de la référence et une copie de l'alternatif ;
    - 1/1 : l'échantillon est homozygote de l'alternatif.
- La qualité du génotype (GQ) qui est basée sur le score Phred.
- La profondeur totale sur l'ensemble des génotypes à une position donnée (DP).
- La profondeur par allèle et par génotype (AD). Alors que la profondeur DP décrit le nombre de fois qu'une base est représentée à un site donné ; la valeur de AD elle est le décompte du nombre de reads portant la référence puis celui de l'alternatif.
- La vraisemblance des 3 génotypes 0/0, 0/1 et 1/1 par rapport à GT. Dans le cas des polymorphismes hétérozygotes, on assigne au génotype le plus probable (GT) une probabilité d'existence de 1. Cette probabilité représentée sur une échelle Phred est utilisée comme base pour établir la vraisemblance des autres génotypes par rapport à ce génotype le plus probable.

Annexe 10 : Principe et fonctionnement de la recalibration



A partir d'un set de SNP source et du jeu de SNP à recalibrer, le VariantRecalibrator compose un jeu commun aux deux précédents. C'est ce noyau commun qui va servir à calculer un seuil de qualité global en deçà duquel les polymorphismes seront considérés comme faux positifs.

Les critères suivant permettent de caractériser un polymorphisme. A chacun est attribué l'ensemble de ces valeurs. VariantRecalibrator permet d'en sélectionner certaines à partir desquelles le VQSLOD sera calculé.

AC : allele count

AF : allele frequency

AN : number of alleles

BaseQRankSum : The u-based z-approximation from the Mann-Whitney Rank Sum Test for base qualities

DP : depth of coverage

DS : Were any of the samples downsampled because of too much coverage?

FS : Phred-scaled p-value using Fisher's Exact Test to detect strand bias (the variation being seen on only the forward or only the reverse strand) in the reads. More bias is indicative of false positive calls

HaplotypeScore : Consistency of the site with two (and only two) segregating haplotypes

InbreedingCoeff : F measures the probability that two genes at any locus in an individual are identical by descent from the common ancestor(s) of the two parents. This means the degree to which two alleles are more likely to be homozygous (AA or aa) rather than heterozygous (Aa) in an individual

MQ : mapping quality (Root Mean Square of the mapping quality of the reads across all samples)

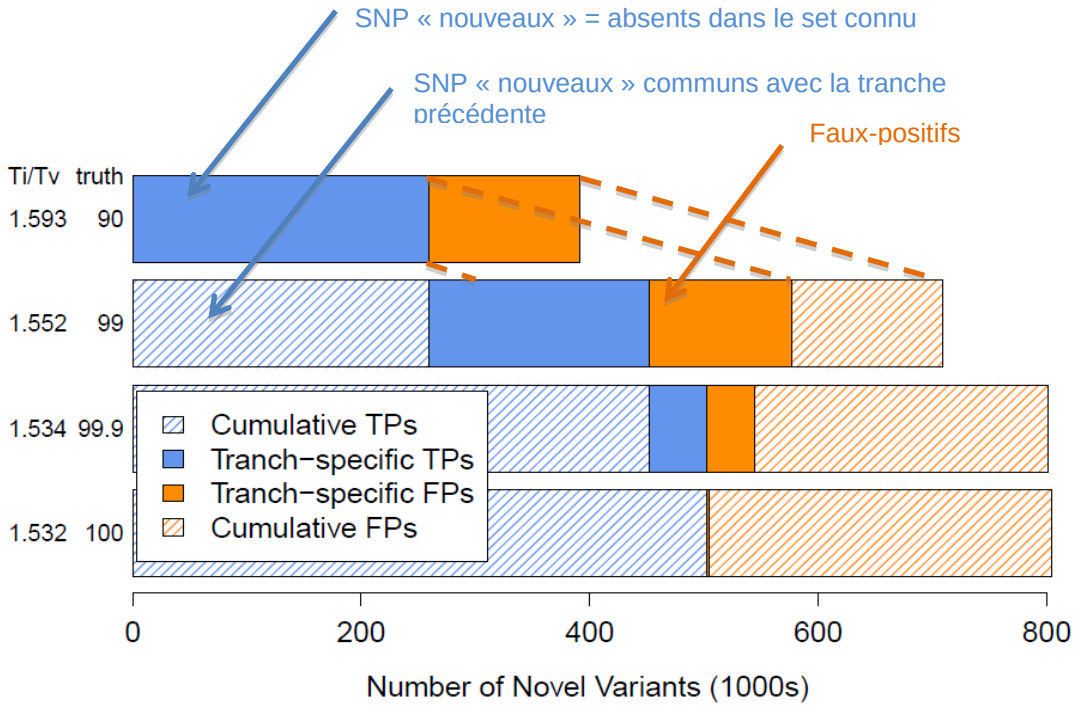
MQ0 : mapping quality zéro (Total count across all samples of mapping quality zero reads)

MQRankSum : The u-based z-approximation from the Mann-Whitney Rank Sum Test for mapping qualities (reads with ref bases vs

QD : Qualitybydepth (Variant confidence (given as (AB+BB)/AA from the PLs) / unfiltered depth)

ReadPosRankSum : The u-based z-approximation from the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele; if the alternate allele is only seen near the ends of reads this is indicative of error)

Pour estimer le taux de faux positif VariantRecalibrator se base sur le Ti/Tv calculé à partir du jeu de SNP sources.



Annexe 11 : Paramètres du Blast

Ci-dessous les paramètres utilisé pour la recherche d'homologies entre les contigs issus de l'assemblage et les banques de biomolécules

program	blastx
database	nr
format	xml
max_target_seq	1
evalue	10
max_thread	24
task	megablast
num_seq_by_batch	8

Annexe 12 : Tableau récapitulatif du nombre de séquences nettoyées, mappées, non mappées.

			Génome			Transcriptome		
	Number of sequences produced	Number of clean sequences	Number of mapped sequences on Btx623	Number of unmapped sequences	Total sequences in mapping	Percentage mapped sequences on Btx623	Number of mapped sequences on Btx623	Number of unmapped sequences	Total sequences in mapping	Percentage mapped sequences on Btx623
EB	74 685 206	59 684 868	12 677 204	8 484 642	21 161 846	0,21	7 481 927,00	29 190 308,00	36 672 235,00	0,13
EC1	13 964 604	11 709 228	8 174 123	2 646 205	10 820 328	0,70	8 665 388,00	2 237 075,00	10 902 463,00	0,74
EC2	14 160 830	11 673 784	8 465 157	2 216 851	10 682 008	0,73	8 397 274,00	2 356 872,00	10 754 146,00	0,72
EC3	17 602 144	14 584 298	9 833 742	3 324 202	13 157 944	0,67	10 283 192,00	2 992 805,00	13 275 997,00	0,71
EC4	32 392 258	25 676 856	17 600 874	6 041 842	23 642 716	0,69	19 383 238,00	4 785 078,00	24 168 316,00	0,75
EC5	17 687 474	14 892 452	10 364 752	3 177 937	13 542 689	0,70	11 150 381,00	2 816 161,00	13 966 542,00	0,75
EC6	15 123 618	13 061 366	9 326 438	2 460 260	11 786 698	0,71	9 526 742,00	2 710 787,00	12 237 529,00	0,73
EC7	17 464 510	14 523 304	10 577 362	3 104 494	13 681 856	0,73	11 263 576,00	2 656 031,00	13 919 607,00	0,78
EC8	18 401 652	15 646 612	11 457 508	3 028 294	14 485 802	0,73	11 999 009,00	2 787 542,00	14 786 551,00	0,77
EC9	12 928 506	11 204 356	8 184 129	2 466 066	10 650 195	0,73	8 914 281,00	1 905 296,00	10 819 577,00	0,80
EC10	15 010 978	12 567 332	9 204 396	2 267 193	11 471 589	0,73	9 540 385,00	2 245 861,00	11 786 246,00	0,76
ES1	40 358 128	30 370 250	8 439 224	2 972 297	11 411 521	0,28	5 516 469,00	8 876 329,00	14 392 798,00	0,18
ES2	34 311 484	25 703 080	8 538 785	1 445 021	9 983 806	0,33	4 932 397,00	9 069 265,00	14 001 662,00	0,19
ES3	31 895 582	24 206 348	14 468 449	3 263 940	17 732 389	0,60	13 011 663,00	5 757 495,00	18 769 158,00	0,54
ES4	44 022 420	34 503 830	23 047 327	6 999 335	30 046 662	0,67	24 417 650,00	6 505 877,00	30 923 527,00	0,71
ES5	33 647 652	25 032 330	9 342 068	1 727 413	11 069 481	0,37	6 306 859,00	6 100 552,00	12 407 411,00	0,25
ES6	22 842 200	16 907 662	10 432 892	4 026 844	14 459 736	0,62	11 673 548,00	2 911 168,00	14 584 716,00	0,69
ES7	27 979 502	21 130 074	13 182 191	4 747 676	17 929 867	0,62	15 194 662,00	3 122 040,00	18 316 702,00	0,72
ES8	41 112 046	31 421 128	21 518 630	6 386 348	27 904 978	0,68	24 191 256,00	4 471 042,00	28 662 298,00	0,77
ES9	41 136 624	32 336 958	17 416 856	5 460 412	22 877 268	0,54	18 648 494,00	6 318 099,00	24 966 593,00	0,58
ES10	44 522 898	34 576 798	22 398 264	5 775 224	28 173 488	0,65	21 961 834,00	7 052 601,00	29 014 435,00	0,64
Total	611	481	264	82	346		262	116	379	

Annexe 13 : Filtres appliqués aux alignements

Le Filter (De Mita, 2012) permet d'éliminer toutes les séquences qui ont trop de données manquantes. Le script se base sur un ratio. Si ce ratio est égal à 0.5, toutes les séquences qui ont moins de 50% de données exploitables de la séquence qui en possède le plus dans l'alignement vont être éliminées. Le filter a été appliqué sur un jeu de séquences déjà filtrées, c'est à dire dont le *nseff*=0 ou *lseff*=0. Pour faciliter l'avancement, le ratio a été arbitrairement établi à 1.0.

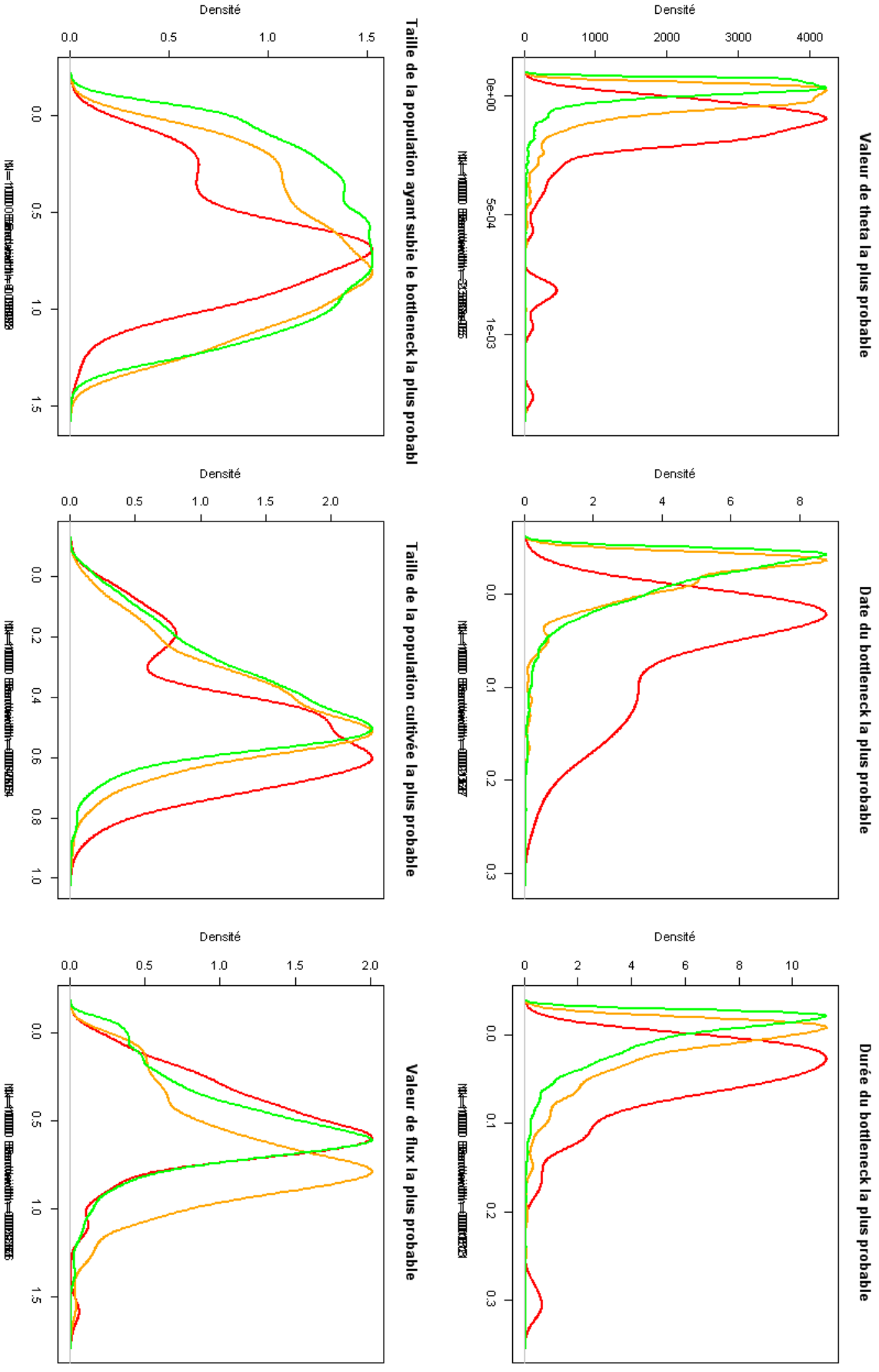
Pour garantir la juste valeur des estimateurs nous avons jugé optimal de conserver les séquences dans lesquelles au moins 100 bases exploitables sont disponibles et les alignements contenant au moins 12 séquences par population (cultivée et sauvage).

Suite à un problème logiciel nous avons pris la décision d'éliminer les séquences propres à l'outgroup, notées @999.

Annexe 14 : Tableau récapitulatif des différents estimateurs en fonction du nombre de simulations

	10000	100000	1000000	
Theta	3,35E-06	2,77E-06	3,43E-07	Min
	6,90E-05	7,10E-05	6,71E-05	Q1
	1,22E-04	1,43E-04	1,24E-04	Medianne
	1,88E-04	1,84E-04	1,68E-04	Moyenne
	1,96E-04	2,24E-04	1,93E-04	Q3
	1,26E-03	2,52E-03	3,29E-03	Max
Size	0,03644	0,001734	0,001401	Min
	0,40833	0,399448	0,407974	Q1
	0,51199	0,537226	0,554643	Medianne
	0,48594	0,511272	0,525759	Moyenne
	0,6243	0,636713	0,659326	Q3
	0,84878	1,04381	1,14928	Max
Date	0,001293	8,15E-05	2,14E-06	Min
	0,015076	3,43E-02	3,03E-02	Q1
	0,038765	8,39E-02	7,13E-02	Medianne
	0,065097	1,19E-01	9,94E-02	Moyenne
	0,105082	1,57E-01	1,37E-01	Q3
	0,248189	1,20E+00	9,30E-01	Max
Dur	0,0003809	1,45E-05	5,22E-07	Min
	0,017361	2,69E-02	3,38E-02	Q1
	0,0412001	6,33E-02	7,74E-02	Medianne
	0,0549626	9,70E-02	1,12E-01	Moyenne
	0,0679349	1,34E-01	1,57E-01	Q3
	0,313768	8,96E-01	1,26E+00	Max
Strength	0,06131	0,000299	0,0008508	Min
	0,44544	0,305888	0,367952	Q1
	0,66557	0,53328	0,644165	Medianne
	0,62658	0,523192	0,6391376	Moyenne
	0,80553	0,737204	0,9071603	Q3
	1,28975	1,06215	1,39529	Max
Migr	0,02114	0,008624	0,006689	Min
	0,38659	0,33651	0,348839	Q1
	0,54402	0,480809	0,48176	Medianne
	0,51499	0,445422	0,458442	Moyenne
	0,64955	0,565488	0,571126	Q3
	1,58286	1,05334	1,34712	Max

Annexe 14bis : Distribution des paramètres en fonction du nombre de simulation. En rouge : 10 000, en orange : 100 000 et en vert : 1 000 000.



Annexe 15 : Tableau récapitulatif des outliers détectés sur la base du ratio de diversité nucléotidique entre le compartiment sauvage et le compartiment cultivé

Type d'Outlier	Gène	Longueur (bp)	Cult_S	Cult_theta	Sauv_S	Sauv_theta	Cult/Sauv_Theta	Annotation fonctionnelle (Phytozome)
Réduction de diversité chez Cultivés	Sb01g041880_1_pacid_1954168	221	0	0,00000	8	0,01020	**0**	Glucan endo-1,3-beta-glucosidase 7
	Sb01g036770_1_pacid_1953538	118	0	0,00000	4	0,01022	**0**	DNAJ/HSP40
	Sb02g000810_2_pacid_1955330	167	0	0,00000	6	0,01028	**0**	Tetratricopeptide repeat(TPR)-containing protein-like
	Sb01g017180_1_pacid_1951380	132	0	0,00000	4	0,01071	**0**	Cytochrome P450 CYP23 subfamily
	Sb01g049250_1_pacid_1955068	522	0	0,00000	20	0,01080	**0**	Histone H4
	Sb10g026090_1_pacid_1984567	235	0	0,00000	9	0,01133	**0**	4 Cl-1 channel CLC-7 and related proteins CLC superfamily
	Sb06g030180_1_pacid_1974068	191	0	0,00000	7	0,01152	**0**	Pas d'annotation fonctionnelle
	Sb10g025880_1_pacid_1984543	113	0	0,00000	3	0,01163	**0**	Putative GDP-L-fucose synthase 2
	Sb09g001250_1_pacid_1979449	295	0	0,00000	8	0,01188	**0**	Replication factor A1, RFA1
	Sb03g013070_1_pacid_1961605	167	0	0,00000	7	0,01219	**0**	Putative pectinacetylesterase
	Sb03g034850_2_pacid_1963322	239	0	0,00000	12	0,01437	**0**	Auxin responsive factor 3
	Sb06g011767_1_pacid_1971998	166	0	0,00000	7	0,01491	**0**	Auxin responsive factor 4
	Sb08g018190_2_pacid_1978619	128	0	0,00000	6	0,01657	**0**	Glucosidase II beta subunit-like protein
	Sb03g034850_1_pacid_1963321	211	0	0,00000	14	0,01870	**0**	Auxin responsive factor 3
	Sb02g028520_1_pacid_1958072	183	0	0,00000	19	0,04547	**0**	Transport transmembranaire
Excès de diversité chez Cultivés	Sb03g046300_1_pacid_1964678	389	7	0,00542	1	0,00091	**5,97**	Pas d'annotation fonctionnelle
	Sb10g000380_1_pacid_1982179	608	6	0,00285	1	0,00047	**6,06**	Pas d'annotation fonctionnelle
	Sb10g001970_1_pacid_1982376	2016	26	0,00364	4	0,00059	**6,19**	Transmembrane protein kinase
	Sb01g028160_1_pacid_1952455	1080	9	0,00262	1	0,00036	**7,34**	HOMEOBOX PROTEIN HOMEOBOX-leucine zipper protein ROC3
	Sb08g018160_1_pacid_1978615	1781	10	0,00186	1	0,00025	**7,56**	DNA replication licensing factor, MCM7 component
	Sb07g018030_1_pacid_1975826	515	20	0,01170	2	0,00150	**7,81**	Pas d'annotation fonctionnelle
	Sb07g022480_1_pacid_1976265	678	8	0,00333	1	0,00042	**7,88**	Pas d'annotation fonctionnelle
	Sb03g010140_1_pacid_1961248	840	8	0,00268	1	0,00034	**8,00**	Putative N-amidino-scyllo-inosamine-4-phosphate phosphatase
	Sb09g002660_1_pacid_1979625	452	7	0,00450	0	0,00000	**Sauv_SEO**	Pas d'annotation fonctionnelle
	Sb07g027370_1_pacid_1976881	207	4	0,00553	0	0,00000	**Sauv_SEO**	40S ribosomal protein S11 Family member
	Sb08g022100_1_pacid_1979144	202	3	0,00419	0	0,00000	**Sauv_SEO**	Amino acid permease family protein, putative, expressed
	Sb02g025246_1_pacid_1957648	127	6	0,01332	0	0,00000	**Sauv_SEO**	CSLC2 (cellulose synthase like)
	Sb01g032850_1_pacid_1953043	332	9	0,00898	0	0,00000	**Sauv_SEO**	MAP kinase-activating protein C22orf5
	Sb06g033430_2_pacid_1974456	138	6	0,01226	0	0,00000	**Sauv_SEO**	Plant phosphoribosyltransferase C-terminal
	Sb07g020440_1_pacid_1976005	225	5	0,00646	0	0,00000	**Sauv_SEO**	Pas d'annotation fonctionnelle

Annexe 16 : Tableau récapitulatif des outliers détectés sur la base de la différenciation entre le compartiment sauvage et le compartiment cultivé

Type d'Outlier	Gène	Longueur (bp)	Global_S	Cult_S	Sauv_S	Kst	Annotation fonctionnelle (Phytozome)
Kst élevé (CI95%)	Sb10g021790_1_pacid_1984022	495	5	0	5	**0,38383**	Phosphatidylethanolamine-binding protein
	Sb01g008695_1_pacid_1950362	359	3	1	3	**0,40240**	PHOSPHATIDYLINOSITOL 3-KINASE TOR1
	Sb07g020170_1_pacid_1975968	1154	26	1	26	**0,40523**	similar to Adenosine 5'-phosphosulfate reductase 5
	Sb03g031860_1_pacid_1962949	294	1	0	1	**0,41353**	similar to DNA-binding protein RAV1-like
	Sb01g005600_1_pacid_1949973	137	2	0	2	**0,41353**	Replication factor C, subunit RFC3
	Sb01g042486_1_pacid_1954238	203	1	1	0	**0,41353**	Pas d'annotation fonctionnelle
	Sb01g040960_1_pacid_1954057	190	1	1	0	**0,41538**	Predicted Ca2+-dependent phospholipid-binding protein
	Sb03g034720_1_pacid_1963305	401	1	1	1	**0,41728**	Pas d'annotation fonctionnelle
	Sb01g042300_1_pacid_1954217	3151	24	19	22	**0,42738**	similar to Nup133 nucleoporin family protein
	Sb07g020260_1_pacid_1975978	916	3	3	2	**0,42844**	Tyrosine kinase specific for activated (GTP-bound) p21cdc42Hs
	Sb01g019560_2_pacid_1951677	101	4	4	4	**0,43101**	Predicted calmodulin-binding protein
	Sb01g008670_1_pacid_1950358	328	3	2	2	**0,43293**	NPH3 family, signal transducer activity, response to light
	Sb03g005050_1_pacid_1960609	532	17	16	16	**0,43553**	RNA BINDING PROTEIN
	Sb01g034790_1_pacid_1953284	221	2	2	1	**0,43834**	Pas d'annotation fonctionnelle
	Sb01g039150_1_pacid_1953838	1502	9	4	8	**0,48166**	Pas d'annotation fonctionnelle
	Sb02g031590_1_pacid_1958442	449	6	6	6	**0,51160**	Protein of unknown function (DUF620)
	Sb01g008680_1_pacid_1950359	408	1	0	1	**0,53333**	similar to Transducin family protein
	Sb06g021260_1_pacid_1973000	339	3	0	2	**0,53571**	similar to Endochitinase A precursor
	Sb10g001710_1_pacid_1982342	815	2	1	1	**0,55119**	pectinesterase activity
	Sb03g035500_1_pacid_1963401	1522	4	0	3	**0,62222**	similar to Probable indole-3-acetic acid-amido synthetase GH3.2
Kst élevé (CI:99%)	Sb07g019816_1_pacid_1975917	300	4	0	2	**0,83909**	HISTONE-LYSINE N-METHYLTRANSFERASE, SUVH
	Sb02g023830_1_pacid_1957471	189	1	0	0	**1,00000**	MAPKK-RELATED SERINE/THREONINE PROTEIN KINASES
Kst faible (CI99.9%)	Sb03g044300_1_pacid_1964440	180	1	1	1	**-0,07714**	Pas d'annotation fonctionnelle
	Sb10g030010_1_pacid_1985069	1475	1	1	1	**-0,07714**	Protein of unknown function (DUF1668)
	Sb08g003850_1_pacid_1977578	226	1	1	1	**-0,07714**	Pas d'annotation fonctionnelle
	Sb10g001590_1_pacid_1982330	112	4	4	3	**-0,06469**	Pas d'annotation fonctionnelle
	Sb02g004470_1_pacid_1955804	290	1	1	1	**-0,06465**	Polyadenylate-binding protein (RRM superfamily)
	Sb02g039810_1_pacid_1959473	196	4	4	4	**-0,06465**	TREHALOSE-6-PHOSPHATE SYNTHASE, similar to Sister of Ramosa 3
	Sb10g029270_1_pacid_1984979	108	1	1	1	**-0,06465**	porphobilinogen synthase [EC:4.2.1.24]
	Sb09g019420_1_pacid_1980757	323	1	1	1	**-0,06009**	3'-5' EXONUCLEASE ERI1-RELATED
	Sb01g009240_1_pacid_1950432	120	1	1	1	**-0,06009**	Domain of unknown function (DUF2828)
	Sb07g027795_1_pacid_1976930	532	1	1	1	**-0,05682**	HYPOXIA-INDUCIBLE FACTOR 1 ALPHA INHIBITOR-RELATED
	Sb04g000807_1_pacid_1964940	464	1	1	1	**-0,05610**	ATP-DEPENDENT HELICASE SMARCA-RELATED
	Sb01g001660_1_pacid_1949484	118	1	1	1	**-0,05610**	Late embryogenesis abundant protein
	Sb04g006475_1_pacid_1965668	750	3	3	3	**-0,05556**	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
	Sb10g002850_1_pacid_1982485	534	1	1	1	**-0,05238**	Early nodulin 93 ENOD93 protein
	Sb10g023660_2_pacid_1984273	140	8	8	8	**-0,05082**	Pas d'annotation fonctionnelle
	Sb01g017900_1_pacid_1951469	288	1	1	1	**-0,05000**	Pas d'annotation fonctionnelle
	Sb04g030210_1_pacid_1967778	234	1	1	1	**-0,05000**	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
	Sb09g030970_1_pacid_1982152	277	1	1	1	**-0,04991**	Pas d'annotation fonctionnelle
	Sb04g015850_1_pacid_1966326	460	1	1	1	**-0,04991**	Pas d'annotation fonctionnelle
	Sb01g044505_1_pacid_1954475	457	1	1	1	**-0,04991**	Pas d'annotation fonctionnelle
	Sb03g046530_1_pacid_1964706	646	1	1	1	**-0,04853**	Protein of unknown function (DUF630)
	Sb04g023430_1_pacid_1966926	216	2	2	2	**-0,04762**	Protein of unknown function (DUF3537)
	Sb03g010900_1_pacid_1961336	414	1	1	1	**-0,04744**	Tetratricopeptide repeat
	Sb04g006763_1_pacid_1965708	119	4	4	3	**-0,04708**	Mitochondrial/chloroplast ribosomal protein L12
	Sb06g032530_1_pacid_1974351	129	4	4	4	**-0,04656**	Pas d'annotation fonctionnelle
	Sb06g018146_1_pacid_1972600	325	3	2	3	**-0,04613**	Protein of unknown function (DUF3537)
	Sb02g025080_1_pacid_1957626	236	2	2	2	**-0,04541**	AP2-like factor, ANT lineage
	Sb0011s012840_1_pacid_1949078	108	1	1	1	**-0,04498**	Pollen proteins Ole e 1 like
	Sb02g005713_1_pacid_1955975	166	1	1	1	**-0,04433**	CENTROMERE-BINDING PROTEIN 1, CBP-1
	Sb01g007970_1_pacid_1950267	747	3	3	3	**-0,04396**	Kinesin-like protein
	Sb03g044720_1_pacid_1964491	842	1	1	1	**-0,04211**	Protein of unknown function (DUF3754)
	Sb03g008420_1_pacid_1961049	546	1	1	1	**-0,04211**	Pas d'annotation fonctionnelle
	Sb02g004530_1_pacid_1955814	416	1	1	1	**-0,04211**	PENTATRICOPEPTIDE REPEAT-CONTAINING PROTEIN
	Sb01g043030_1_pacid_1954307	1030	1	1	1	**-0,04042**	beta-glucosidase [EC:3.2.1.21]
	Sb08g023067_1_pacid_1979282	350	1	1	1	**-0,03963**	Phosphoribulokinase / Uridine kinase family
	Sb03g009000_1_pacid_1961114	393	1	1	1	**-0,03947**	Protein involved in vacuolar protein sorting
	Sb06g025850_1_pacid_1973544	515	17	17	17	**-0,03871**	RING FINGER PROTEIN 24-RELATED, Zinc finger, C3HC4 type (RING finger)
	Sb07g028156_1_pacid_1976974	600	1	1	1	**-0,03639**	Pas d'annotation fonctionnelle
	Sb07g028945_1_pacid_1977069	194	1	1	1	**-0,03639**	CENTROMERE-BINDING PROTEIN 1, CBP-1
	Sb10g005920_1_pacid_1982877	301	1	1	1	**-0,03597**	UDP-glucuronate 4-epimerase [EC:5.1.3.6]
	Sb03g041430_1_pacid_1964098	308	1	1	1	**-0,03534**	response to stress
	Sb04g007940_2_pacid_1965860	122	1	1	1	**-0,03518**	PROTEIN PHOSPHATASE PP2A REGULATORY SUBUNIT B

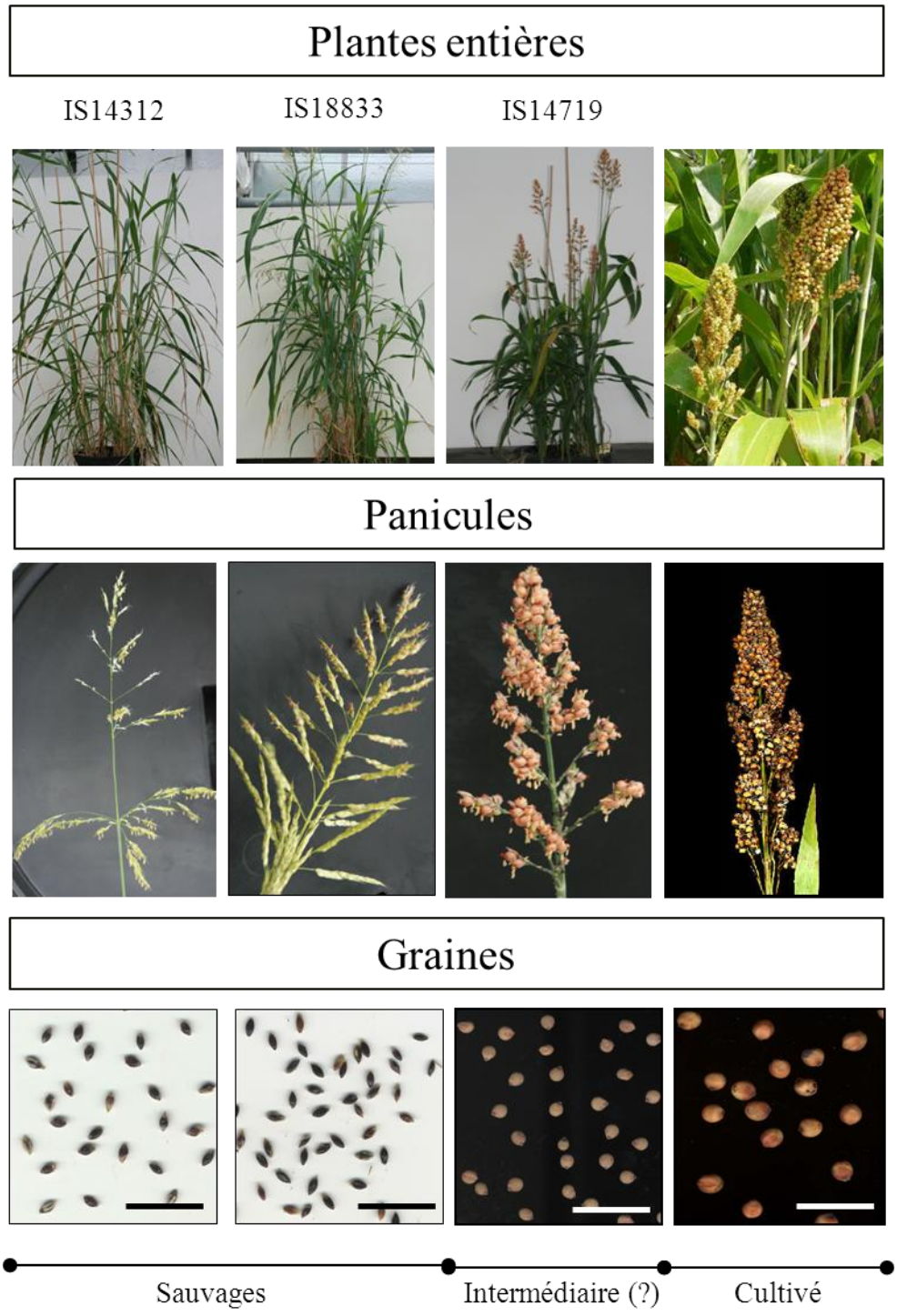
Annexe 17 : Tableau récapitulatif des outliers détectés sur la base du test D de Tajima au sein du compartiment cultivé

Type d'Outlier	Gène	Longueur (bp)	Cult_S	Cult_Tajima_D	Sauv_S	Sauv_Tajima_D	Annotation fonctionnelle (Phytozome)
D de Tajima élevé chez cultivés	Sb08g017420_1_pacid_1978528	433	7	-3,39	2	-1,55	Pas d'annotation fonctionnelle
	Sb07g024910_1_pacid_1976574	1233	40	-2,57	48	-1,58	Centromere-associated protein HEC1
	Sb04g027920_1_pacid_1967489	1497	8	-2,17	22	2,76	similar to Chromosome structural maintenance protein-like
	Sb01g040750_1_pacid_1954036	2461	23	-2,15	35	-1,84	similar to Beta-galactosidase 6 precursor, BETA-GALACTOSIDASE RELATED
	Sb03g025230_1_pacid_1962106	1735	7	-2,12	11	0,65	cell communication, protein binding, phosphoinositide binding
	Sb07g026870_1_pacid_1976822	276	7	-2,12	9	-2,24	similar to Oxysterol-binding protein-like
	Sb07g004890_2_pacid_1975188	695	9	-2,10	17	1,54	mRNA export protein, contains WD40 repeats)
	Sb03g009560_1_pacid_1961181	668	31	-2,07	28	0,73	similar to MFS18 protein precursor
	Sb02g030180_1_pacid_1958280	458	6	-2,05	8	-1,28	Pas d'annotation fonctionnelle
	Sb01g040950_1_pacid_1954056	1078	5	-1,97	6	-0,31	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
	Sb01g047830_1_pacid_1954898	1624	12	-1,97	16	-0,69	Protein phosphatase 2C/pyruvate dehydrogenase (lipoamide) phosphatase
	Sb03g038260_1_pacid_1963731	1481	14	-1,96	21	0,40	Endoplasmic reticulum vesicle transporter
	Sb02g027070_1_pacid_1957895	1252	24	-1,93	31	0,09	phosphoribosylaminoimidazole-succinocarboxamide synthase [EC:6.3.2.6]
	Sb03g026160_1_pacid_1962232	2995	8	-1,91	20	-0,92	HEAT repeat
	Sb10g025520_1_pacid_1984501	1413	8	-1,91	6	-1,55	Serine carboxypeptidase S28
D de Tajima faible chez cultivés	Sb05g018936_1_pacid_1970307	2650	34	3,02	37	-0,04	Pas d'annotation fonctionnelle
	Sb05g019560_1_pacid_1970374	666	10	3,03	10	1,39	protein-phycocyanobilin linkage, CpeT/CpcT family (DUF1001)
	Sb04g034590_1_pacid_1968303	2146	34	3,04	34	3,12	TETRATRICOPEPTIDE-LIKE HELICAL
	Sb07g028660_1_pacid_1977035	2035	54	3,04	54	-0,58	similar to Putative signal transducer and activator of transcription interacting protein
	Sb03g007100_1_pacid_1960870	1276	39	3,04	42	2,44	Pas d'annotation fonctionnelle
	Sb01g008430_1_pacid_1950329	2319	20	3,05	23	1,59	MAPKK-RELATED SERINE/THREONINE PROTEIN KINASES
	Sb07g020350_1_pacid_1975991	376	18	3,08	19	2,71	Protein of unknown function (DUF3353)
	Sb01g018570_1_pacid_1951552	1177	15	3,11	15	1,02	DNA replication factor CDT1 like
	Sb02g038560_1_pacid_1959309	254	14	3,12	15	2,25	similar to Glucosyltransferase S10a-like, UDP-glucuronosyl and UDP-glucosyl transferase
	Sb08g019120_1_pacid_1978748	2311	26	3,12	32	0,73	similar to Diacylglycerol kinase 1, putative, expressed, diacylglycerol kinase
	Sb08g006680_1_pacid_1977924	602	14	3,17	15	2,09	similar to Glutathione S-transferase GSTT18, Glutathione S-transferase
	Sb01g033670_1_pacid_1953154	1154	25	3,19	30	2,02	similar to Nodulin-like protein, Gamma-like transporter family
	Sb02g020320_1_pacid_1957074	1693	23	3,20	25	1,68	similar to Ferrochelatase, FERROCHELATASE
	Sb09g024790_1_pacid_1981409	888	24	3,21	27	2,26	Pas d'annotation fonctionnelle

Annexe 18 : Tableau récapitulatif des outliers détectés sur la base du test D de Tajima au sein du compartiment sauvage

Type d'Outlier	Gène	Longueur (bp)	Cult_S	Cult_Tajima_D	Sauv_S	Sauv_Tajima_D	Function_Phytozome_DP
D de Tajima élevé chez sauvages	Sb10g002600_1_pacid_1982453	442	8	0,86184204	8	-2,4691436	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
	Sb04g025090_1_pacid_1967133	1548	27	1,85904206	26	-2,4556306	similar to 26S proteasome non-ATPase regulatory subunit 5
	Sb04g034260_1_pacid_1968266	1394	3	-1,4405932	16	-2,4524957	similar to Mitochondrial phosphate transporter
	Sb01g019280_1_pacid_1951638	1393	2	-0,7683224	20	-2,4511209	similar to Malate dehydrogenase, cytoplasmic
	Sb02g022770_1_pacid_1957335	1217	0	Cult_S=0	19	-2,4393262	similar to Probable protein arginine N-methyltransferase 1
	Sb10g027690_1_pacid_1984776	1986	1	0,35109766	15	-2,4245896	similar to Putative uncharacterized external alternative NADPH dehydrogenase
	Sb03g014460_1_pacid_1961747	3481	1	1,02588309	18	-2,4243451	similar to Putative uncharacterized protein HASTY
	Sb08g006220_1_pacid_1977864	1917	1	0,35195258	18	-2,4242537	similar to Leafbladeless1
	Sb10g029170_1_pacid_1984966	1030	2	-0,7685046	22	-2,4146142	similar to 3-phosphoinositide-dependent protein kinase-1-like
	Sb08g012560_1_pacid_1978176	10996	14	-0,9118762	63	-2,4137867	Domain of Unknown Function (DUF913)
	Sb01g019420_1_pacid_1951656	1249	1	-0,5915097	17	-2,4130498	Uncharacterized conserved protein
	Sb09g024840_1_pacid_1981414	2019	3	-0,8754998	21	-2,4045382	similar to Ferredoxin-sulfite reductase precursor
	Sb01g035380_1_pacid_1953367	1531	2	-0,7678053	16	-2,3968325	UDP-glucose 4-epimerase [EC:5.1.3.2]
	Sb04g020560_1_pacid_1966576	1558	2	-0,7684918	20	-2,3911433	ubiquitin-like modifier-activating enzyme 5
	Sb01g018130_1_pacid_1951497	3786	9	-1,2485631	24	-2,3871724	NA
D de Tajima faible chez sauvages	Sb10g030030_1_pacid_1985072	573	8	0,21677303	7	2,84195213	NADH-UBIQUINONE OXIDOREDUCTASE B22 SUBUNIT
	Sb01g049210_1_pacid_1955063	885	12	-0,6401892	15	2,84576487	Pas3 annotation nonfunctionnelle
	Sb10g026530_1_pacid_1984618	371	12	0,79229371	12	2,84966204	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
	Sb01g038660_1_pacid_1953776	1215	23	-0,6482171	24	2,86527477	similar to Phosphatidylinositol transfer protein, expressed
	Sb06g021200_2_pacid_1972994	220	10	0,91585914	9	2,87493551	Pas3 annotation nonfunctionnelle
	Sb06g029910_1_pacid_1974036	521	3	0,84866309	12	2,87819029	Pas3 annotation nonfunctionnelle
	Sb01g048100_1_pacid_1954929	299	11	2,44962866	10	2,88018962	Similar to Chitin elicitor-binding protein precursor
	Sb04g034830_1_pacid_1968335	4249	33	2,80871925	33	2,88134864	similar to Putative myosin heavy chain
	Sb03g045040_1_pacid_1964530	286	8	1,21393349	8	2,88193103	PsbP-related thylakoid lumenal protein 2;
	Sb09g004830_1_pacid_1979904	569	16	1,85251393	15	2,92853221	oxidoreductase activity, acting on the CH-CH group of donors
	Sb09g026150_1_pacid_1981569	1648	50	0,68105717	52	2,96780172	PHOSPHOFRUCTOKINASE
	Sb05g000790_1_pacid_1968875	489	9	2,98230113	9	2,97310053	Double-stranded RNA binding motif
	Sb01g010830_1_pacid_1950642	1276	21	2,7171852	21	3,03072457	similar to Beta-glucosidase
	Sb04g034590_1_pacid_1968303	2146	34	3,03898431	34	3,11785135	TETRATRICOPEPTIDE-LIKE HELICAL

Annexe 19 : Illustrations d'accessions étudiées au cours du projet, mise en évidence du caractère atypique d'IS14719



Résumé :

L'optimisation de la gestion de la diversité et de l'efficacité de la sélection repose sur une meilleure connaissance des facteurs génétiques et des forces évolutives affectant la variabilité des caractères d'intérêt agronomique et adaptatif. En parallèle des approches de génétique quantitative et moléculaire, une approche évolutive basée sur l'analyse de la diversité nucléotidique permet d'identifier les régions portant des signatures des évènements de domestication ou d'adaptation à des contraintes naturelles.

En définissant le scénario le plus probable d'évolution du sorgho à l'aide de données moléculaires issues de la portion exprimée du génome d'un échantillon de 20 individus représentatifs de la diversité mondiale, on prend en compte l'influence des phénomènes démographiques sur l'évolution neutre des gènes ce qui permet d'identifier ceux présentant des profils de diversité divergents des attendus neutres.

Le modèle de domestication estimé à partir de 200 000 polymorphismes répartis sur environ 15 000 gènes permet d'approximer la date de domestication du sorgho à 5600 ans avant le présent, résultat en accord avec les autres études sur le sujet. Le modèle sur lequel on s'appuie, permet également l'estimation de l'expansion de la population ayant subi le goulot d'étranglement de force et de durée en accord avec les données de la littérature. La reconstruction de l'histoire évolutive du sorgho nous a permis d'identifier les gènes potentiellement impliqués dans le processus de domestication.

Ainsi, grâce à l'analyse de l'indice de différenciation Kst entre les compartiments sauvage et cultivé, 64 gènes présentant des différenciations extrêmes par rapport aux attendus neutres ont été détectés. Parmi eux, on trouve des gènes intervenant dans le contrôle de la croissance cellulaire, de la réponse aux nutriments et aux facteurs de croissance et dans la réponse aux stress.

Mots-clés : Diversité génétique, sorgho, domestication, approche ABC, SNP/indels

Abstract:

Optimization of genetic diversity management in terms of conservation and breeding purposes relies on an in-depth understanding of the genetic factors and evolutionary forces acting on the variability of the traits of adaptive and agronomic interest. In synergy with the quantitative and molecular genetic approaches, evolutionary approaches based on the analysis of patterns of nucleotide diversity allows identifying genomic regions harboring molecular signatures of domestication and adaptation.

Through the definition of the most likely history of cultivated sorghum based on sequence diversity information from 20 accessions representatives of the sorghum worldwide diversity, demographic events which have affected the evolution of sorghum genes are taken into account to identify non neutral genes (i.e. domestication or adaptive genes).

The domestication model developed, based on 200 000 polymorphisms identified on around 15 000 genes, allowed to estimate a domestication event around 5600 years BP, this result being in accordance with previous estimations. Bottleneck strength, importance of cultivated population expansion and bidirectional migration rate were also evaluated and allowed to reconstruct sorghum history. This information was then used to identify genes potentially involved in the domestication process or of adaptive importance.

As an example, the analysis of the differentiation between the cultivated and wild pools allowed the identification of 64 genes harboring non neutral diversity patterns. It is interesting to note that a significant part of them is involved in cellular growth control, responses to nutrient availability, growth factors or abiotic and biotic stress.

Keywords : Genetic diversity, sorghum, domestication, ABC approach, SNP/indels